SEPLN
SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL

ISSN: 1135-5948

Artículos

Tesis

ISSN: 1135-5948

Comité Editorial

ISSN: 1135-5948
ISSN electrónico: 1989-7553
Depósito Legal: B:3941-91
Editado en: Universidad de Alicante
Año de edición: 2024
Editores:

Eugenio Martínez Cámara	Universidad de Jaén	emcamara@ujaen.es
Álvaro Rodrigo Yuste	UNED	alvarory@lsi.uned.es
Aitziber Atutxa Salazar	Universidad del País Vasco	aitziber.atucha@ehu.eus

Publicado por: Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, EdificioA3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

SEPLN
SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL

ISSN: 1135-5948

Preámbulo

La revista *Procesamiento del Lenguaje Natural* pretende ser un foro de publicación de artículos científico-técnicos inéditos de calidad relevante en el ámbito del Procesamiento de Lenguaje Natural (PLN) tanto para la comunidad científica nacional e internacional, como para las empresas del sector. Además, se quiere potenciar el desarrollo de las diferentes áreas relacionadas con el PLN, mejorar la divulgación de las investigaciones que se llevan a cabo, identificar las futuras directrices de la investigación básica y mostrar las posibilidades reales de aplicación en este campo. Anualmente la SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural) publica dos números de la revista, que incluyen artículos originales, reseñas bibliográficas, resúmenes de tesis doctorales y resúmenes de las tareas del Foro de Evaluación de Lenguas Ibéricas (IberLEF). Esta revista se distribuye gratuitamente a todos los socios, y con el fin de conseguir una mayor expansión y facilitar el acceso a la publicación, su contenido es libremente accesible por Internet.

Las áreas temáticas tratadas son las siguientes:

- Modelos lingüísticos, matemáticos y psicolingüísticos del lenguaje
- Lingüística de corpus
- Desarrollo de recursos y herramientas lingüísticas
- Gramáticas y formalismos para el análisis morfológico y sintáctico
- Semántica, pragmática y discurso
- Lexicografía y terminología computacional
- Resolución de la ambigüedad léxica
- Aprendizaje automático en PLN
- Generación textual monolingüe y multilingüe
- Traducción automática
- Reconocimiento y síntesis del habla
- Extracción y recuperación de información monolingüe, multilingüe y multimodal
- Sistemas de búsqueda de respuestas
- Análisis automático del contenido textual
- Resumen automático
- PLN para la generación de recursos educativos
- PLN para lenguas con recursos limitados
- Aplicaciones industriales del PLN
- Sistemas de diálogo
- Análisis de sentimientos y opiniones
- Minería de texto
- Evaluación de sistemas de PLN
- Implicación textual y paráfrasis

El ejemplar número 72 de la revista *Procesamiento del Lenguaje Natural* contiene trabajos correspondientes a comunicaciones científicas y resúmenes de tesis doctorales. Todos ellos han sido aceptados mediante el proceso de revisión tradicional en la revista. Queremos agradecer a los miembros del Comité Asesor y a los revisores adicionales la labor que han realizado.

Se recibieron 33 trabajos para este número, de los cuales 30 eran artículos científicos y 3 resúmenes de tesis doctorales. De entre los 30 artículos recibidos, 10 han sido finalmente seleccionados para su publicación, lo cual fija una tasa de aceptación del 33,33%.

El Comité Asesor de la revista se ha hecho cargo de la revisión de los trabajos. Este proceso de revisión es de doble anonimato: se mantiene oculta la identidad de los autores que son evaluados y de los revisores que realizan las evaluaciones. En un primer paso, cada artículo ha sido examinado de manera ciega o anónima por tres revisores. En un segundo paso, para aquellos artículos que tenían una divergencia mínima de tres puntos (sobre siete) en sus puntuaciones, sus tres revisores han reconsiderado su evaluación en conjunto. Finalmente, la evaluación de aquellos artículos que estaban en posición muy cercana a la frontera de aceptación ha sido supervisada por más miembros del comité editorial. El criterio de corte adoptado ha sido la media de las tres calificaciones, siempre y cuando hayan sido iguales o superiores a 5 sobre 7.

Marzo de 2024
Los editores.

ISSN: 1135-5948

Preamble

The *Natural Language Processing* journal aims to be a forum for the publication of high-quality unpublished scientific and technical papers on Natural Language Processing (NLP) for both the national and international scientific community and companies. Furthermore, we want to strengthen the development of different areas related to NLP, widening the dissemination of research carried out, identifying the future directions of basic research and demonstrating the possibilities of its application in this field. Every year, the Spanish Society for Natural Language Processing (SEPLN) publishes two issues of the journal that include original articles, book reviews, summaries of doctoral theses and summaries of the shared tasks of the Iberian Languages Evaluation Forum (IberLEF). All issues published are freely distributed to all members, and contents are freely available online.

The subject areas addressed are the following:

- Linguistic, Mathematical and Psychological models to language
- Grammars and Formalisms for Morphological and Syntactic Analysis
- Semantics, Pragmatics and Discourse
- Computational Lexicography and Terminology
- Linguistic resources and tools
- Corpus Linguistics
- Speech Recognition and Synthesis
- Dialogue Systems
- Machine Translation
- Word Sense Disambiguation
- Machine Learning in NLP
- Monolingual and multilingual Text Generation
- Information Extraction and Information Retrieval
- Question Answering
- Automatic Text Analysis
- Automatic Summarization
- NLP Resources for Learning
- NLP for languages with limited resources
- Business Applications of NLP
- Sentiment Analysis
- Opinion Mining
- Text Mining
- Evaluation of NLP systems
- Textual Entailment and Paraphrases

The 72nd issue of the *Procesamiento del Lenguaje Natural* journal contains scientific papers and doctoral dissertation summaries. All of these were accepted by a peer review process. We would like to thank the Advisory Committee members and additional reviewers for their work.

Thirty-three papers were submitted for this issue, from which thirty were scientific papers and three doctoral dissertation summaries. From these thirty papers, we selected ten (33,33%) for publication.

The Advisory Committee of the journal has reviewed the papers in a double-blind process. Under double-blind review the identity of the reviewers and the authors are hidden from each other. In the first step, each paper was reviewed blindly by three reviewers. In the second step, the three reviewers have given a second overall evaluation of those papers with a difference of three or more points out of seven in their individual reviewer scores. Finally, the evaluation of those papers that were in a position very close to the acceptance limit were supervised by the editorial board. The cut-off criterion adopted was the mean of the three scores given.

March 2024
Editorial board.

ISSN: 1135-5948

Artículos

Tesis

Información General

Artículos

Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 15-30 recibido 18-11-2023 revisado 10-01-2024 aceptado 30-01-2024

Revisiting Challenges and Hazards in Large Language Model Evaluation

Análisis de los Desafíos y Riesgos en la Evaluación de Grandes Modelos del Lenguaje

Inigo Lopez-Gazpio
HiTZ Basque Center for Language Technology - Ixa NLP Group
University of the Basque Country UPV/EHU
inigo.lopez@ehu.eus

Abstract: In the age of large language models, artificial intelligence's goal has evolved to assist humans in unprecedented ways. As LLMs integrate into society, the need for comprehensive evaluations increases. These systems' real-world acceptance depends on their knowledge, reasoning, and argumentation abilities. However, inconsistent standards across domains complicate evaluations, making it hard to compare models and understand their pros and cons. Our study focuses on illuminating the evaluation processes for these models. We examine recent research, tracking current trends to ensure evaluation methods match the field's rapid progress requirements. We analyze key evaluation dimensions, aiming to deeply understand factors affecting models performance. A key aspect of our work is identifying and compiling major performance challenges and hazards in evaluation, an area not extensively explored yet. This approach is necessary for recognizing the potential and limitations of these AI systems in various domains of the evaluation.
Keywords: Large language models, evaluation, evaluation challenges and hazards, evaluation dimensions.

Resumen: En la era de los modelos de lenguaje de gran escala, el objetivo de la inteligencia artificial ha evolucionado para asistir a personas de maneras sin precedentes conocidos. A medida que los modelos se integran en la sociedad, aumenta la necesidad de evaluaciones exhaustivas. La aceptación de estos sistemas en el mundo real depende de sus habilidades de conocimiento, razonamiento y argumentación. Sin embargo, estándares inconsistentes entre dominios complican la evaluación, dificultando la comparación de modelos y la comprensión de su funcionamiento. Nuestro estudio se enfoca en organizar y aclarar los procesos de evaluación de estos modelos. Examinamos investigaciones recientes para analizar las tendencias actuales e investigar si los métodos de evaluación se ajustan a los requisitos del progreso. Finalmente, identificamos y detallamos los principales desafíos y riesgos que afectan la evaluación, un área que aún no ha sido explorada extensamente. Este enfoque es necesario para reconocer las limitaciones actuales, el potencial y las particularidades de la evaluación de estos sistemas.
Palabras clave: Modelos de lenguaje de gran escala, evaluación, desafíos y riesgos de evaluación, dimensiones de la evaluación.

1 Introduction

Since the early days of expert systems, it has been recognized that for these systems to be accepted in real-world domains, they must not only demonstrate their knowledge (Khalfa, 1994) but also be able to reason and argue about it (Buchanan and Shortliffe, 1984; Lacave and Díez, 2002; Korb and Nicholson, 2010). In the era of large language models (LLM), the aim of artificial intelligence has shifted from merely imitating natural intelligence to supporting humans in novel unprecedented ways (Deng and Lin, 2022). The acceptance of AI by users hinges on the quality of the evaluations performed. The advent of pre-trained language models has marked a significant advancement. These models, developed by training Transformer models (Vaswani et al., 2017) on extensive corpora, have exhibited exceptional capabilities in various

ISSN 1135-5948 DOI 10.26342/2024-72-1

natural language processing (NLP) tasks, sometimes presumably surpassing human performance (Orrù et al., 2023; Hadi et al., 2023a; Zhao et al., 2023; Chang et al., 2023). The recent surge in LLM performance evaluation reflects the complexity and necessity of tailored evaluation approaches.

Recently, the evaluation of LLMs has continuously evolved, placing greater focus on evaluation beyond fixed knowledge traditional datasets and focusing on innovative aspects, such as: comprehensive assessments (Xu et al., 2023a), ethical considerations (Head et al., 2023), and sustainability (Khowaja, Khuwaja, and Dev, 2023). These aspects are now considered alongside the traditional evaluations of knowledge and generalization capabilities. As LLMs increasingly become a part of our societal frameworks, the need for multidimensional and thorough evaluations becomes more pronounced. This diverse range of evaluation approaches not only improves the quality of LLMs but also ensures their responsible and advantageous application in real-world scenarios. Consequently, the process of evaluating LLMs has become a crucial component closely tied to the development and refinement of these models.

The evaluation criteria for LLMs, including BERT (Aftan and Shah, 2023), GPT-3 (Floridi and Chiriatti, 2020), InstructGPT (Ouyang et al., 2022), PaLM (Chowdhery et al., 2022), GPT-4 (OpenAI, 2023), LLaMA (Touvron et al., 2023) and their successors (Lehman et al., 2023), is turning into a complex and multifaceted process crucial for understanding their capabilities, limitations, and impacts. Traditional metrics like perplexity (Gamallo, Campos, and Alegria, 2017) and BLEU score (Reiter, 2018), focusing on linguistic accuracy and fluency, are no longer sufficient (Tang, Chuang, and Hu, 2023). As LLMs become more advanced, their evaluation also needs to evolve, encompassing a broader range of criteria to ensure their robustness, effectiveness, fairness, interpretability, environmental impact and safety in task-specific settings. Recent evaluations have concentrated on several key aspects:

1. Robustness and Generalization: testing across diverse topics and contexts to ensure consistent performance even in unfamiliar scenarios. Generalization tests are essential as they assess a model's ability to effectively apply its acquired knowledge to new and unfamiliar domains (Dong et al., 2023).

2. Fairness and Bias Testing: LLMs can inadvertently perpetuate societal biases present in their training data. Rigorous testing is required to identify and mitigate biases to prevent discrimination over race, gender, or other sensitive attributes (Li et al., 2023; Huang et al., 2023).

3. Interpretability and Explainability: understanding the decision-making process of LLMs is vital. Interpretability tools and methods are being developed to provide insights into model's knowledge. Transparency is crucial for trust and reliability in sensitive applications (Saha et al., 2023).

4. Environmental Impact: computational demands of training and operating large models have brought attention to their environmental effects. Assessing these models for energy efficiency and carbon footprint is now crucial, guiding the field towards sustainable practices (Rillig et al., 2023).

5. Task-Specific Evaluations: task-specific evaluations are vital beyond just general metrics. For instance, in a translation task, fluency and cultural appropriateness are key, while in a medical diagnosis application, accuracy and reliability are paramount (Chang et al., 2023).

6. Human-Centric Evaluations: including human judgment in evaluation processes is becoming more popular. Human evaluators offer detailed feedback on elements such as usefulness, coherence, empathy, and the suitability of responses, areas where automated metrics may fall short (Ouyang et al., 2022; Zhong et al., 2023).

7. Adversarial Testing: Exposing LLMs to adversarial examples, where inputs are deliberately modified to test the model's resilience, is another emerging evaluation strategy. This helps in understanding the limits of a model's understanding and reasoning capabilities (Xu et al., 2023b).

Recent advancements in LLM evaluations have led to diverse, non-standardized approaches. A comprehensive evaluation approach is crucial for developing robust, fair, and efficient models, but it introduces challenges such as complexity, consistency, resource demands, and adaptability. The wide array of evaluation areas requires unique methodologies, tools, and expertise, making the process complex and resource-intensive. With varying standards and benchmarks across domains, consistency in evaluations is difficult, complicating model comparisons and full understanding of their strengths and weaknesses. Furthermore, finding a balance among different evaluation criteria is challenging. Enhancing a model's performance in one task could compromise its effectiveness in another area. Under these circumstances, there's also a risk of overemphasizing certain domains, like reading comprehension, machine translation or generability, at the expense of others, such as truthfulness, fairness or interpretability. This imbalance can lead to models excelling in certain tasks but falling short in vital areas. Additionally, domains involving human-centric criteria, such as ethics or user satisfaction, bring subjectivity into evaluations, causing inconsistent outcomes and interpretations. For newcomers or smaller institutions, the broad spectrum of evaluation areas poses a challenge. The need for extensive resources and expertise to perform thorough evaluations may limit innovation and diversity in the research community

In the current landscape, where even the challenges of LLM evaluation are not clearly defined, this study aims to clarify the well-known evaluation domains of LLMs. By reviewing the latest in LLM research, we aim to highlight recent trends and keep evaluation methods aligned with the rapid developments in this field. An ongoing challenge is to ensure these evaluation domains and methodologies remain updated. Additionally, this research attempts to link the main hazards associated with key LLM evaluation dimensions, an area that has not yet been thoroughly explored. Understanding these hazards is crucial for creating more effective evaluation scenarios for LLMs. However, tackling these hazards demands a multi-disciplinary approach that goes beyond technical solutions, incorporating considerations of ethics, user experience, and societal impact. As LLMs continue to advance, methods for evaluating and addressing these hazards must also evolve.

This study is organized as follows: Section 1, "Introduction", sets the stage and context for our work. Section 2, "Review on LLM evaluation", reviews the dimensions of LLM evaluation based on current research and analyzes the performance of state-of-the-art LLMs. Section 3, "Discussion on LLM evaluation", delves into a detailed discussion on LLM evaluation, highlighting the primary hazards associated with these evaluation dimensions. Section 4, "Description of main hazards", specifically focuses on identifying and detailing the main hazards in LLM evaluation. Finally, Section 5, "Conclusions", summarizes our findings, outlines future research directions, and discusses the limitations.

2 Review on LLM evaluation

LLMs are increasingly popular in both academic and industrial settings due to their remarkable performance across various applications (Devlin et al., 2018; Gao and Lin, 2004; Kasneci et al., 2023; Zhao et al., 2023). As LLMs become more integral to research and everyday use, understanding their potential risks at both task and societal levels is essential. Recent years have seen considerable efforts in evaluating and assessing LLMs from multiple angles. Typically, LLMs are defined as language models with hundreds of billions of parameters, trained on vast text datasets (Shanahan, 2022). Most LLMs share similar model architectures, based in the Transformer architecture, and pre-training objectives, such as language modeling, with size variable training parameters. The key distinction of LLMs lies in their significantly larger scale in terms of model size, data used for training, and computational power. This scaling enables them to better comprehend natural language and generate high-quality text based on given contexts or prompts. The improvement in capability with model size is partially explained by the scaling law, where performance increases substantially with model size (Kaplan et al., 2020). However, certain abilities, as noted in (Zhao et al., 2023), only become apparent when the model size reaches a specific threshold, deviating from what the scaling law predicts.

LLMs have recently received substantial interest in both academic and industrial sectors (Bommasani et al., 2021; Wei et al., 2022; Zhao et al., 2023). As indicated by recent research (Bubeck et al., 2023), the impressive performance of LLMs has sparked optimism about their potential as a form of Artificial General Intelligence (AGI). Unlike previous models limited to specific tasks, LLMs are adept at a wide range of tasks, from general language tasks to domain-specific applications. This versatility makes them increasingly popular among users with critical information needs.

Furthermore, these billion-parameter models, despite being resource-intensive, are surprisingly user-friendly. They don't demand access to specialized hardware or software, nor a deep understanding of machine learning or natural language processing. Instead, LLMs are accessible through APIs and are capable –or at least claimed to be– of handling complex tasks with minimal (few-shot) or no (zero-shot) prior information. This accessibility offers a more intuitive and natural way of interacting with computers (de Wynter et al., 2023). The complexity inherent in the linguistic interactions of a LLM makes it challenging to establish a concise, standardized method for assessing its quality or gaining a deeper understanding of how to evaluate its composed representations. Consequently, a diverse array of evaluation methods for LLMs is emerging to address these multiple challenges.

This section provides a detailed review of the principal methods used to evaluate LLMs in the state-of-the-art, highlighting several critical dimensions. In line with recent trends, these evaluations focus on various aspects: (i) robustness and generalization reliability of the models, (ii) fairness and the presence of bias in model outputs, (iii) interpretability and explainability of the models, (iv) environmental impact of the models, (v) task-specific evaluation such as translation or summarization, (vi) human-centric evaluation including user trust and confidence, and (vii) resilience against adversarial testing.

Current consensus in the field of LLM evaluation suggests that it should be structured around three key dimensions, each encompassing distinct aspects and challenges: (dimension 1) the scope of the evaluation, (dimension 2) the extent of the evaluation, and (dimension 3) the procedure of the evaluation.

The studies conducted by (Orrù et al., 2023; Hadi et al., 2023a; Chang et al., 2023; Zhao et al., 2023) are among the first comprehensive surveys in this area. They concur on the importance of these three dimensions for LLM evaluation. The first dimension covers the range of evaluation tasks applicable to LLMs. The second dimension focuses on selecting suitable scenarios for the evaluation (i.e. benchmarks). The third dimension deals with the actual evaluation process, employing the chosen tasks, datasets or benchmarks. These dimensions collectively form the cornerstone of effective evaluation. We will now describe each of these dimensions in detail.

2.1 Fixed-knowledge evaluation

Evaluating fixed-knowledge in LLMs is complex, with no universal solution fitting all scenarios. The primary aim of such evaluations is to compare different systems that generate varied representations for a specific task. Although the ultimate objective is to apply these models in high-level tasks or market applications, evaluating them on manually annotated, more detailed tasks often provides deeper insights and facilitates error analysis in controlled environments. In fact, focusing on intermediate tasks has been instrumental in advancing fixed-knowledge evaluation, thanks to widely-used datasets in key natural language processing (NLP) categories.

Main categories in NLP encompass Natural Language Understanding (Bates, 1995) and Natural Language Generation (McDonald, 2010) tasks. Examples include text classification (Song et al., 2014), reading comprehension (Baradaran, Ghiasi, and Amirkhani, 2022), machine translation (Baltrušaitis, Ahuja, and Morency, 2018), language modeling (Min et al., 2023), grammar analysis (Wang et al., 2020), code generation (Shin and Nam, 2021), question answering (Bouziane et al., 2015), dialogue (Motger, Franch, and Marco, 2022), logic reasoning (Costantini, 2002), language inference (Storks, Gao, and Chai, 2019), truthfulness (Oshikawa, Qian, and Wang, 2018), fact checking (Lazarski, Al-Khassaweneh, and Howard, 2021), toxicity detection (Garg et al., 2023), bias detection (Garg et al., 2023), multimodality (Erdem et al., 2022), summarization (Awasthi et al., 2021), negation (Mahany et al., 2022), sentiment analysis (Zhang, Wang, and Liu, 2018), semantic understanding (Salloum, Khan, and Shaalan, 2020), and more.

2.2 Evaluation of versatility

Evaluating how well foundational models handle tasks at a human level is crucial in their development towards AGI. Traditional fixed-knowledge datasets, often based on single tasks might not fully capture human-like abilities, as the latter ones potentially combine multiple objectives.Thus, the approach of fixed-knowledge evaluation for LLMs is becoming recognized as inadequate for a thorough assessment. This method, which uses a static set of datasets, falls short due to the dynamic and complex nature of language and knowledge, as well as the continuous evolution of

LLMs. Such evaluations don't always reflect the real-world versatility and adaptability required of these advanced systems.

The shortcomings of fixed-knowledge evaluation have prompted the creation of large-scale, dynamic benchmarks. These benchmarks are tailored to encompass a wider range of language understanding and generation tasks, striving to be more inclusive and reflective of real-world language usage. They typically involve diverse and complex tasks, extensive enough to capture broad linguistic trends. Additionally, these benchmarks often incorporate considerations of fairness, bias detection, and ethics, acknowledging the increasing importance of social responsibility in LLMs. By assessing models against these expanded criteria, we can better ensure their linguistic proficiency as well as their ethical and social integrity (Zhong et al., 2023).

Recently, a variety of benchmarks have been developed to evaluate LLMs across a range of tasks. We now enumerate and briefly describe some of the most notables:

GLUE (Wang et al., 2018) (General Language Understanding Evaluation) and SUPERGLUE (Wang et al., 2018) consist of sota benchmarks designed to mimic real-world language processing scenarios. They encompass a variety of tasks such as text classification, machine translation, reading comprehension, and dialogue generation, offering a comprehensive assessment of capabilities.

PromptBench (Zhu et al., 2023) highlights the sensitivity of current LLMs to adversarial prompts, underscoring the need for meticulous prompt engineering.

WinoGrande (Sakaguchi et al., 2021) is a benchmark designed to test AI systems' common sense reasoning and natural language understanding. It features a series of nearly identical sentence pairs, each with a subtle variation that alters the meaning of a crucial word. The test for AI systems is to accurately interpret these sentences and resolve the ambiguities.

AGIEVAL (Zhong et al., 2023) stands out as a human-centric benchmark based on standardized exams. It encompasses a diverse array of tests, including college entrance exams, law school admission tests, math competitions, and lawyer qualification exams. This benchmark is designed to evaluate AI systems in contexts that require a high level of academic and professional understanding.

Another significant benchmark is MMLU (Hendrycks et al., 2020) (Massive Multitask Language Understanding). MMLU offers a comprehensive evaluation framework to test AI models' language understanding across various subjects and disciplines. It includes tasks from humanities and social sciences to STEM fields, aiming to gauge the models' depth and breadth of knowledge. MMLU is distinctive for its focus on complex comprehension and reasoning, challenging language models to demonstrate their understanding and processing abilities across diverse areas of expertise.

BigBench (Ghazal et al., 2013) is recognized as an industry-standard benchmark for big data analytics. BIG-bench benchmark serves as a thorough and varied tool for evaluating LLMs. It covers a broad spectrum of tasks, testing different aspects of NLU and NLG, and extends beyond the scope of traditional benchmarks. BIG-bench is specifically designed to challenge LLMs in areas like advanced reasoning, creativity, and comprehension of complex and subtle language nuances.

HELM (Liang et al., 2022) offers a comprehensive evaluation framework for LLMs. It assesses language models on multiple fronts, including NLU, NLG, coherence, context sensitivity, common-sense reasoning, and domain-specific knowledge. The goal of HELM is to provide a holistic evaluation of language models, gauging their performance across a variety of tasks and domains.

HellaSwag (Zellers et al., 2019) is a benchmark specifically designed to assess common sense reasoning and contextual understanding in LLMs. It provides context-rich scenarios, each accompanied by multiple-choice endings, and the model's task is to select the most plausible conclusion for each scenario. The scenarios in HellaSwag are intentionally diverse and challenging, often demanding a nuanced comprehension of everyday activities and situations. This benchmark aims to advance AI capabilities in complex, real-world common sense reasoning.

The HumanEval benchmark (Chen et al., 2021) is designed to test the code generation abilities of LLMs. It presents a series of programming challenges, each consisting of a function signature, a body with a TODO comment, and several unit tests. The model's task is to complete the function body so that it successfully passes all the tests. HumanEval specifically focuses on models' capacity for understanding and generating functional programming code. It evaluates the algorithmic thinking, problem-solving, and coding skills, making it an important tool for gauging software development skills.

The GSM benchmark (Cobbe et al., 2021) is tailored to test LLMs' mathematical reasoning skills. It comprises grade-school level math problems that span a range of mathematical skills, from basic arithmetic to advanced problem-solving. This benchmark challenges AI models to comprehend and manipulate numerical information, execute calculations, and utilize mathematical concepts to solve problems. GSM is particularly valuable for evaluating capabilities in logical reasoning and numerical understanding.

2.3 Methodology of the evaluation

The third dimension of evaluation revolves around the evaluation methodology and particularly whether human judgment is incorporated into the process. Incorporating human feedback into the evaluation of LLMs is becoming increasingly essential, complementing automated scoring metrics like BLEU or perplexity. While automated metrics offer valuable quantitative data, they often miss the nuanced, qualitative elements of language crucial for a comprehensive understanding and enhancement of knowledge-based systems (Qin et al., 2023; Bang et al., 2023).

Automated metrics are typically designed to assess specific linguistic aspects, such as grammatical accuracy or lexical similarity to a reference text. However, effective language use involves more than just grammatical correctness. It encompasses context, cultural nuances, pragmatics, and the conveyance of subtle meanings, which automated metrics may not fully grasp. Human evaluators bring a crucial perspective to these qualitative elements, providing a more complete evaluation of performance. Furthermore, human evaluation is key in determining the relevance and coherence of LLM-generated content (Novikova et al., 2017). A model may generate text that scores highly on automated metrics like BLEU or perplexity, but this doesn't guarantee that the content is contextually appropriate or coherent. Human reviewers are able to assess if the text is useful, logical and consistent within its context, factually accurate, and maintains overall coherence.

Another crucial aspect of LLM evaluation is assessing creativity and novelty in language use. As LLMs are increasingly employed for creative tasks the limitations of automated metrics become evident (Bubeck et al., 2023). These metrics typically rely on comparisons with existing data and are not equipped to judge originality. Human evaluators, on the other hand, can appreciate and assess creativity, offering insights vital for fostering innovation in model development. Moreover, human input is indispensable in detecting and addressing biases in LLM outputs. Automated metrics fall short in identifying biases or ethical concerns in generated content. Human evaluators, with their understanding of societal and cultural nuances, are better positioned to spot when a model outputs biased or potentially harmful content. This human oversight is crucial for the development of responsible and ethical AI systems. Additionally, human evaluators play a pivotal role in user experience testing, particularly for LLM applications designed for human interaction (Demetriadis and Dimitriadis, 2023). Human feedback on the engagement, usefulness, and enjoyment level of these interactions is invaluable, as it provides insights that automated metrics cannot capture. This human-in-the-loop approach ensures that the models are not only technically proficient but also effective and satisfying in real-world interactions.

2.4 Qualitative performance

Much of the leading research on LLM evaluation involves empirical assessments using many well-known models (Xu et al., 2022; Lai et al., 2023; de Wynter et al., 2023; Zhao et al., 2023; Zhang et al., 2023; Koh, Salakhutdinov, and Fried, 2023; Liu et al., 2021). This includes GPT-3, GPT-3.5, InstructGPT, LlaMa, PaLM, and their variants. This subsection synthesizes findings from readily available off-the-shelf models and public research or leaderboard results. The goal is to summarize the overall qualitative performance of LLMs as reflected in current state-of-the-art. Notable evaluations of LLMs are detailed in studies like (Hadi et al., 2023a; Zhao et al., 2023). These investigations assess the effectiveness and superiority of LLMs across a broad range of tasks and benchmarks, particularly in relation to the first and second dimensions of evaluation defined in Section 2.

Regarding the first dimension of evaluation, (Zhao et al., 2023) primarily focused on language generation tasks, including language modeling, conditional text generation, and code synthesis. They also concentrated on knowledge utilization and complex reasoning tasks. The authors aimed to cover the most widely discussed or studied tasks in LLM evaluation, rather than encompassing all specific tasks in the NLU and NLG fields. The findings from this investigation align with those of (Brown et al., 2020; Costa-jussà et al., 2022), showing that LLMs significantly outperform previous state-of-the-art methods on fixed-knowledge evaluation datasets. This is evident in public leaderboards (e.g., SNLI, MNLI matched, MNLI mismatched, X-NLI), where LLMs with billions of parameters demonstrate clear superiority over smaller models in considerable sized fixed-knowledge datasets. (Kaplan et al., 2020) noted that performance in language modeling tasks tends to adhere to the scaling law. This suggests that increasing the size of language models leads to improved accuracy and lower perplexity, further underscoring the advantages of scaling up LLMs.

Conditional text generation, a key task in NLG, focuses on creating text that meets specific requirements based on given conditions. Studies by (Li et al., 2022; Zhao et al., 2023) identify conditional generation as a complex task, requiring at least an understanding of machine translation, text summarization, and question answering. While evaluation for these tasks often intersects with the second dimension of evaluation, involving the use of segments from various fixed-knowledge datasets to create more intricate benchmarks, LLMs have shown exceptional performance. They excel not only on existing datasets but also on these comprehensive benchmarks, in some cases even presumably outperforming human abilities due to their advanced language generation skills. In line with these developments, (OpenAI, 2023) reported significant progress with GPT-4. This model has presumably already surpassed state-of-the-art methods, including those with benchmark-specific training, across a broad array of tasks like NLU, commonsense reasoning, and mathematical reasoning. Yet, the true nature of the model is not known, nor the evaluation procedures employed in the validation of the model. There might be several factors that affect the cited surpass, such as data contamination among others. As a consequence, until the evaluation process is clarified it must be doubted that the nature of that surpass is due to the model generalization capabilities and not to contamination (Sainz et al., 2023).

The study by (Bubeck et al., 2023) goes a step further by likening GPT-4 to an early form of AGI. They highlight GPT-4's human-like performance in real-world exams such as Advanced Placement tests and the Graduate Record Examination, covering areas like mathematics, computer vision, and programming. However, they also note significant limitations in GPT-4's performance. Consistent with the scaling law observed in fixed-knowledge evaluation, GPT-4 shows marked improvements over GPT-3.5, which itself surpassed earlier GPT versions. (Bang et al., 2023) provide a detailed analysis in which they demonstrate (in 9 out of 13 NLP datasets) the superiority of modern GPT over earlier LLMs using zero-shot learning. Their work also reveals that recent GPT versions outdo fully fine-tuned task-specific language models in 4 different tasks on the MMLU benchmark. For the rest of the scenarios, GPT's performance is comparable to, or slightly below, that of fully fine-tuned models, though statistical significance in these comparisons is not always clear. (Srivastava et al., 2022) corroborate these findings. They show that GPT-3 with context can surpass a fine-tuned BERT-Large on SuperGLUE score with only 32 example inputs. This further substantiates the scaling law's impact on LLM performance, which is still in need of further investigation. In their analysis of MMLU, (Hoffmann et al., 2022) demonstrate that LLMs nearly double the average accuracy of human raters. Notably, GPT-4 exhibits state-of-the-art performance in 5-shot settings, achieving an average accuracy improvement of over 10 % compared to the previously best-performing model.

Regarding the third dimension of evaluation, comparisons and investigations involving LLMs are less common, partly due to the high costs and complexities involved. However, recent studies, including (Creswell, Shanahan, and Higgins, 2022), indicate that automatic metrics might underestimate the quality of LLM-generated content, while human judgment tends to offer more favorable assessments. This finding outlines the increasing necessity of incorporating human evaluation into the loop, highlighting its crucial role in providing a more accurate measure of LLMs' generation capability and quality.

As efforts continue to focus on the development of new metrics that better align with human judgment, human-in-the-loop LLM evaluation is increasingly incorporating tasks that mimic pseudo-human judgments, like code synthesis. In this task, LLMs are required to do more than just generate high-quality natural language as they also need to demonstrate proficiency in creating formal language that meets specific human-defined conditions (Wang et al., 2022). This shift not only tests LLMs' natural language abilities but also their capability to adhere to structured coding requirements, offering a more comprehensive evaluation framework.

Unlike in NLG, the quality of the generated code can be directly verified through execution with appropriate compilers or interpreters. Current research performed in this domain often evaluates the effectiveness of LLMs by measuring the pass rate of the generated code against human-designed test cases, lending this method a pseudo-human evaluation character. Recent developments have seen the introduction of several code benchmarks focused on functional correctness to assess LLMs' code synthesis capabilities. As these tasks increase in complexity, smaller models often perform almost as random baselines (Perez et al., 2022; Bradbury et al., 2018; Nijkamp et al., 2023).

3 Discussion on LLM evaluation

Overall, state-of-the-art results in LLM evaluation reveal that increasing model size seems to continuously improve performance. (Chowdhery et al., 2022) report that the most advanced LLMs can surpass average human performance in many scenarios under a few-shot setting, particularly in well-known benchmarks assessing the models' generalizing capabilities across various fixed knowledge settings. It is important to recognize that human performance by itself is not universally defined within the state of the art. This variability underscores the complexity of directly comparing LLM capabilities with human benchmarks. However, understanding when and how LLMs develop these abilities is crucial, as highlighted by (Fu, Peng, and Khot, 2022). The fact that LLMs are primarily developed by industry players, who often don't disclose critical training details like data collection and cleaning, complicates efforts to replicate and conduct detailed analyses.

(Zhao et al., 2023) argue that, despite their progress and impact, the fundamental mechanisms underlying LLMs remain largely unexplored. Also, there is a notable uncertainty on why highly advanced abilities emerge in LLMs, while the very same abilities are absent in smaller models. This lack of understanding calls for a more in-depth examination of the key factors contributing to the superior capabilities of billion-parameter LLMs.

The study by (Bang et al., 2023) highlights certain drawbacks and limitations of LLMs, particularly in how they generalize. They identify areas where LLMs, specifically GPT variants, struggle. For example, GPT models show weaknesses in inductive reasoning, as opposed to deductive or abductive reasoning. They also lack spatial reasoning capabilities, although they perform better in temporal reasoning. Another significant limitation noted is in mathematical reasoning, a concern also echoed by (Frieder et al., 2023). Furthermore, (Bang et al., 2023) claims that GPT-like models demonstrate acceptable performance in causal and analogical reasoning. They also note that these models are relatively more proficient in commonsense reasoning compared to non-textual semantic reasoning.

All in all, there seems to be an unknown number of hazards affecting the performance of LLMs across different evaluation dimensions, but there has been limited analysis identifying these hazards. (Ji et al., 2023) conducted a thorough investigation into the hallucination hazard, which is one of the most common one. Hallucinations in LLMs refer to factual statements generated by the model that cannot be verified based on the information contained within its parametric memory, spanning all the model's knowledge. While hallucination is perhaps the most recognized hazard associated with LLMs, there exists a range of other, less-known hazards that impact evaluation. These hazards raise critical questions about the effectiveness of existing benchmarks in properly evaluating and reflecting LLMs' capabilities. Acknowledging this challenge, the next section of our study aims to highlight what we consider the most significant performance affecting hazards in LLM evaluation. This analysis spans across the three main dimensions of evaluation, aiming to provide a comprehensive understanding of the factors that influence LLM performance.

4 *Description of main hazards*

This section enumerates a comprehensive list of hazards in LLM evaluation, each linked to a specific area or dimension of evaluation they are associated with. With the understanding of the factors that influence poor performance we aim to clarify the current challenges in each evaluation dimension.

The Reversal Curse. This Natural Language Inference hazard refers to the phenomenon where models incorrectly assign higher probability to the reverse of a true statement. For instance, if a model recognizes "A implies B", it might also incorrectly assess "B implies A" as true, showcasing a fundamental misunderstanding of logical inference (Ma et al., 2023; Berglund et al., 2023).

Lack of Common Sense Reasoning. As a generalization of the previous hazard, LLMs sometimes fail in tasks requiring common sense reasoning, generating outputs that are logically absurd or factually incorrect (Kejriwal et al., 2023).

Hallucination. A content generation domain hazard in which LLMs produce plausible but entirely fabricated information, known as hallucinations. This is particularly hazardous in domains where factual accuracy is critical, such as for fixed-knowledge evaluation. This hazard has been very well documented in the state-of-the-art (Ji et al., 2023; Puchert et al., 2023; Bang et al., 2023).

Interpretability and Explainability Issues. As LLMs grow in complexity, understanding the reasoning behind their decisions becomes more challenging. This lack of transparency is a hazard in applications where understanding model decision-making is crucial for trust and reliability (Saha et al., 2023; Saha et al., 2023).

Catastrophic Forgetting. In the domain of the learning stability, this refers to a model's tendency to forget previously learned information upon learning new data (Zhai et al., 2023; Sun et al., 2020).

Bias and Stereotyping. Linked with fairness and ethics, this hazard states that LLMs can inherit and amplify biases present in their training data. This includes gender, racial, and cultural biases, leading to unfair or stereotypical outputs. This is a significant hazard where fairness and ethical considerations are paramount, such as in the third dimension (Kotek, Dockum, and Sun, 2023).

Model Overfitting and memorization. Generalization hazard that occurs when a model is too closely tailored to the training data and fails to perform well on unseen data (Peng, Wang, and Deng, 2023). This is a critical hazard in evaluating the model's ability to generalize beyond its training set affecting all dimensions. Serious concerns regarding memorization are raised by the authors in (Sainz et al., 2023) where they expose test data from benchmarks being present as training for LLMs in different conditions.

Adversarial Attacks. LLMs can be vulnerable to adversarial attacks affecting the robustness domain, where slight, often imperceptible, alterations to input data can lead to drastically different outputs. This hazard challenges the robustness and security of models (Sainz et al., 2023; Sakaguchi et al., 2021; Xu et al., 2023b).

Vulnerability to Misinformation When trained on data containing misinformation, LLMs can inadvertently propagate false or misleading information (Saha et al., 2023).

Inconsistency in Long-Term Interactions. Similar to the previous hazard, in applications involving long-term interactions, LLMs may exhibit inconsistency in personality or knowledge over time, affecting user experience and trust. Also, LLMs may struggle with understanding and maintaining context over longer conversations or texts, leading to responses that are out of context or irrelevant (Chen, Arunasalam, and Celik, 2023).

Output Toxicity. Affecting content safety domain, LLMs can generate harmful or offensive content, especially if they are exposed to such content. This is a significant hazard in public-facing applications (Chetnani, 2023).

Echo Chamber Effect. In the domain of content diversity LLMs can reinforce the same ideas or perspectives, especially if trained on homogeneous data, leading to a lack of diversity in generated content and potentially reinforcing biases (Demarco, de Zarate, and Feuerstein, 2023).

Language and Cultural Limitations. For cross-lingual domains, LLMs often struggle with languages with low digital resources or with cultural nuances, leading to poor performance in multilingual or multicultural contexts (Hadi et al., 2023b).

Misalignment with Human Values. Concerning the third dimension of evaluation and the ethical alignment, LLMs might generate outputs that are technically correct but misaligned with human ethical standards, especially in sensitive areas like medical, legal, or moral advice. Also, interactions with users can create feedback loops where the model increasingly reinforces user biases or undesirable behaviors (Chiang and Lee, 2023).

Difficulty with Nuanced or Subtle Language. Related to the previous hazard, LLMs may struggle with understanding and generating nuanced or subtle language, such as sarcasm, irony, or metaphor (Băroiu and Trăuşan-Matu, 2023).

Environmental issues. This hazard relates to addressing not only global sustainability goals, but also the long-term viability and ethical development of AI technologies (Rillig et al., 2023). The environmental impact of LLMs emerges as a critical hazard, characterized by the significant energy consumption and carbon footprint associated with their training, validation and operation. This domain transversal hazard underscores the need for sustainability in AI practices, advocating for the development and adoption of energy-efficient algorithms.

Privacy and Copyright. Privacy and copyright issues present a significant hazard in the context of LLMs, reflecting concerns around the unauthorized use of proprietary data and the potential for privacy breaches. Aligning with the OECD's principles on artificial intelligence fairness and ethics, it's crucial to ensure that LLMs operate within frameworks that respect copyright laws and protect personal data.

Benchmark over-reliance. Linked with human-centric evaluation weaknesses, benchmark over-reliance emphasizes the transversal risk of overvaluing benchmark results when assessing NLU capabilities of LLMs. This hazard challenges the notion of superhuman performance, arguing that benchmarks may not fully capture the nuances of human language comprehension and often lack transparency and fairness in comparisons. This hazard calls for the development of more comprehensive and equitable benchmarks to accurately measure and understand the capabilities of language models in relation to human performance (Tedeschi et al., 2023).

5 *Conclusions*

This work provides a comprehensive understanding of the challenges in evaluating LLMs, focusing on the identification of key performance hazards. It emphasizes the need for continuous evolution of evaluation methods to keep up with the advancements in LLM technology and ensure responsible development and deployment. As LLMs become integral to societal frameworks, there is a growing need to emphasize the importance of multidimensional and comprehensive evaluations. The acceptance of these systems in real-world applications is tied not only to their knowledge demonstration but also to their reasoning and argumentation abilities.

Our study reviews recent research in LLMs, tracking current trends to ensure that evaluation methods keep pace with rapid advancements in the field. We analyze key evaluation dimensions with the aim of understanding factors that affect the performance of LLMs. A significant aspect of this investigation is identifying major performance hazards in LLM evaluation, an area not extensively explored previously. This approach is crucial for recognizing the potential and limitations of these AI systems in various evaluation domains. Evaluating LLMs is crucial for several reasons. First, it allows us to understand their strengths and weaknesses more clearly, and, second, enhanced evaluations offer better guidance for human-LLM interactions, informing future interaction designs and implementations.

5.1 Limitations on LLM evaluation

As LLMs grow in size and develop more emergent abilities, current evaluation protocols may no longer suffice to accurately assess their capabilities and potential risks. Therefore, our goal is to heighten awareness within the community about the significance of LLM evaluation. We achieve this by reviewing existing evaluation protocols and, more importantly, by highlighting the need for future research focused on developing new LLM evaluation protocols that take into account the underlying hazards that affect each dimension. This approach is crucial for keeping pace with the rapid advancements in LLM technology and ensuring their responsible development and deployment.

References

Aftan, S. and H. Shah. 2023. A survey on bert and its applications. In *2023 20th Learning and Technology Conference (L&T)*, pages 161–166. IEEE.

Aiyappa, R., J. An, H. Kwak, and Y.-Y. Ahn. 2023. Can we trust the evaluation on chatgpt? *arXiv preprint arXiv:2303.12767*.

Awasthi, I., K. Gupta, P. S. Bhogal, S. S. Anand, and P. K. Soni. 2021. Natural language processing (nlp) based text summarization-a survey. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pages 1310–1317. IEEE.

Baltrušaitis, T., C. Ahuja, and L.-P. Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Bang, Y., S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Baradaran, R., R. Ghiasi, and H. Amirkhani. 2022. A survey on machine reading comprehension systems. *Natural Language Engineering*, 28(6):683–732.

Băroiu, A.-C. and Ş. Trăuşan-Matu. 2023. How capable are state-of-the-art language models to cope with sarcasm? In *2023 24th International Conference on Control Systems and Computer Science (CSCS)*, pages 399–402. IEEE.

Bates, M. 1995. Models of natural language understanding. *Proceedings of the National Academy of Sciences*, 92(22):9977–9982.

Berglund, L., M. Tong, M. Kaufmann, M. Balesni, A. C. Stickland, T. Korbak, and O. Evans. 2023. The reversal curse: Llms trained on.ª is b"fail to learn"b is a". *arXiv preprint arXiv:2309.12288*.

Bommasani, R., D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bouziane, A., D. Bouchiha, N. Doumi, and M. Malki. 2015. Question answering systems: survey and trends. *Procedia Computer Science*, 73:366–375.

Bradbury, J., R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, et al. 2018. Jax: Composable transformations of python+ numpy programs (v0. 2.5). *Software available from https://github. com/google/jax*.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Buchanan, B. G. and E. H. Shortliffe. 1984. *Rule based expert systems: the mycin experiments of the stanford heuristic programming project (the Addison-Wesley series in artificial intelligence)*. Addison-Wesley Longman Publishing Co., Inc.

Chang, Y., X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Chen, M., J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chen, Y., A. Arunasalam, and Z. B. Celik. 2023. Can large language models provide security & privacy advice? measuring the ability of llms to refute misconceptions. *arXiv preprint arXiv:2310.02431*.

Chetnani, Y. P. 2023. *Evaluating the Impact of Model Size on Toxicity and Stereotyping in Generative LLM*. Ph.D. thesis, State University of New York at Buffalo.

Chiang, C.-H. and H.-y. Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Chollet, F. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Chowdhery, A., S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Clark, E., S. Rijhwani, S. Gehrmann, J. Maynez, R. Aharoni, V. Nikolaev, T. Sellam, A. Siddhant, D. Das, and A. P. Parikh. 2023. Seahorse: A multilingual, multifaceted dataset for summarization evaluation. *arXiv preprint arXiv:2305.13194*.

Cobbe, K., V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Costa-jussà, M. R., J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Costantini, S. 2002. Meta-reasoning: A survey. In *Computational Logic: Logic Programming and Beyond: Essays in Honour of Robert A. Kowalski Part II*. Springer, pages 253–288.

Creswell, A., M. Shanahan, and I. Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

de Wynter, A., X. Wang, A. Sokolov, Q. Gu, and S.-Q. Chen. 2023. An evaluation on large language model outputs: Discourse and memorization. *arXiv preprint arXiv:2304.08637*.

Demarco, F., J. M. O. de Zarate, and E. Feuerstein. 2023. Measuring ideological spectrum through nlp. In *Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023) co-located with 22th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2023)*.

Demetriadis, S. and Y. Dimitriadis. 2023. Conversational agents and language models that learn from human dialogues to support design thinking. In *International Conference on Intelligent Tutoring Systems*, pages 691–700. Springer.

Deng, J. and Y. Lin. 2022. The benefits and challenges of chatgpt: An overview. *Frontiers in Computing and Intelligent Systems*, 2(2):81–83.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, G., J. Zhao, T. Hui, D. Guo, W. Wang, B. Feng, Y. Qiu, Z. Gongque, K. He, Z. Wang, et al. 2023. Revisit input perturbation problems for llms: A unified robustness evaluation framework for noisy slot filling task. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 682–694. Springer.

Erdem, E., M. Kuyu, S. Yagcioglu, A. Frank, L. Parcalabescu, B. Plank, A. Babii, O. Turuta, A. Erdem, I. Calixto, et al. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207.

Floridi, L. and M. Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Frieder, S., L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, and J. Berner. 2023. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*.

Fu, Y., H. Peng, and T. Khot. 2022. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. *Yao Fu's Notion*.

Gamallo, P., J. R. P. Campos, and I. Alegria. 2017. A perplexity-based method for similar languages discrimination. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)*, pages 109–114.

Gao, J. and C.-Y. Lin. 2004. Introduction to the special issue on statistical language modeling. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(2):87–93.

Garg, T., S. Masud, T. Suresh, and T. Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s):1–32.

Ghazal, A., T. Rabl, M. Hu, F. Raab, M. Poess, A. Crolotte, and H.-A. Jacobsen. 2013. Bigbench: Towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD international conference on Management of data*, pages 1197–1208.

Hadi, M. U., R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili. 2023a. A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*.

Hadi, M. U., R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, et al. 2023b. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Head, C. B., P. Jasper, M. McConnachie, L. Raftree, and G. Higdon. 2023. Large language model applications for evaluation: Opportunities and ethical implications. *New Directions for Evaluation*, 2023(178-179):33–46.

Hendrycks, D., C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hoffmann, J., S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Huang, D., Q. Bu, J. Zhang, X. Xie, J. Chen, and H. Cui. 2023. Bias assessment and mitigation in llm-based code generation. *arXiv preprint arXiv:2309.14345*.

Jain, N., K. Saifullah, Y. Wen, J. Kirchenbauer, M. Shu, A. Saha, M. Goldblum, J. Geiping, and T. Goldstein. 2023. Bring your own data! self-supervised evaluation for large language models. *arXiv preprint arXiv:2306.13651*.

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jin, Z., J. Liu, Z. Lyu, S. Poff, M. Sachan, R. Mihalcea, M. Diab, and B. Schölkopf. 2023. Can large language models infer causation from correlation? *arXiv preprint arXiv:2306.05836*.

Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kasneci, E., K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Kejriwal, M., H. Santos, K. Shen, A. M. Mulvehill, and D. L. McGuinness. 2023. Context-rich evaluation of machine common sense. In *International Conference on Artificial General Intelligence*, pages 167–176. Springer.

Khalfa, J. 1994. *What is intelligence?* Cambridge University Press.

Khowaja, S. A., P. Khuwaja, and K. Dev. 2023. Chatgpt needs spade (sustainability, privacy, digital divide, and ethics) evaluation: A review. *arXiv preprint arXiv:2305.03123*.

Koh, J. Y., R. Salakhutdinov, and D. Fried. 2023. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*.

Korb, K. B. and A. E. Nicholson. 2010. *Bayesian artificial intelligence*. CRC press.

Kotek, H., R. Dockum, and D. Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pages 12–24.

Lacave, C. and F. J. Díez. 2002. A review of explanation methods for bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127.

Lai, V. D., N. T. Ngo, A. P. B. Veyseh, H. Man, F. Dernoncourt, T. Bui, and T. H. Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Lazarski, E., M. Al-Khassaweneh, and C. Howard. 2021. Using nlp for fact checking: A survey. *Designs*, 5(3):42.

Lehman, J., J. Gordon, S. Jain, K. Ndousse, C. Yeh, and K. O. Stanley. 2023. Evolution through large models. In *Handbook of Evolutionary Machine Learning*. Springer, pages 331–366.

Li, J., T. Tang, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. 2022. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2201.05273*.

Li, Y., M. Du, R. Song, X. Wang, and Y. Wang. 2023. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Liang, P., R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Lin, S., J. Hilton, and O. Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Liu, F., E. Bugliarello, E. M. Ponti, S. Reddy, N. Collier, and D. Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*.

Ma, J.-Y., J.-C. Gu, Z.-H. Ling, Q. Liu, and C. Liu. 2023. Untying the reversal curse via bidirectional language model editing. *arXiv preprint arXiv:2310.10322*.

Mahany, A., H. Khaled, N. S. Elmitwally, N. Aljohani, and S. Ghoniemy. 2022. Negation and speculation in nlp: A survey, corpora, methods, and applications. *Applied Sciences*, 12(10):5209.

McDonald, D. D. 2010. Natural language generation. *Handbook of natural language processing*, 2:121–144.

Min, B., H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Motger, Q., X. Franch, and J. Marco. 2022. Software-based dialogue systems: survey, taxonomy, and challenges. *ACM Computing Surveys*, 55(5):1–42.

Nijkamp, E., H. Hayashi, C. Xiong, S. Savarese, and Y. Zhou. 2023. Codegen2: Lessons for training llms on programming and natural languages. *arXiv preprint arXiv:2305.02309*.

Novikova, J., O. Dušek, A. C. Curry, and V. Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

OpenAI, R. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2.

Orrù, G., A. Piarulli, C. Conversano, and A. Gemignani. 2023. Human-like problem-solving abilities in large language models using chatgpt. *Frontiers in Artificial Intelligence*, 6:1199350.

Oshikawa, R., J. Qian, and W. Y. Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Peng, Z., Z. Wang, and D. Deng. 2023. Near-duplicate sequence search at scale for large language model memorization evaluation. *Proceedings of the ACM on Management of Data*, 1(2):1–18.

Perez, E., S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*.

Puchert, P., P. Poonam, C. van Onzenoodt, and T. Ropinski. 2023. Llmmaps–a visual metaphor for stratified evaluation of large language models. *arXiv preprint arXiv:2304.00457*.

Qin, C., A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476.*

Reiter, E. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Rillig, M. C., M. Ågerstrand, M. Bi, K. A. Gould, and U. Sauerland. 2023. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466.

Ruder, S., J. H. Clark, A. Gutkin, M. Kale, M. Ma, M. Nicosia, S. Rijhwani, P. Riley, J.-M. A. Sarr, X. Wang, et al. 2023. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages. *arXiv preprint arXiv:2305.11938.*

Saha, T., D. Ganguly, S. Saha, and P. Mitra. 2023. Workshop on large language models' interpretability and trustworthiness (llmit). In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5290–5293.

Sainz, O., J. A. Campos, I. García-Ferrero, J. Etxaniz, O. L. de Lacalle, and E. Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018.*

Sakaguchi, K., R. L. Bras, C. Bhagavatula, and Y. Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Salloum, S. A., R. Khan, and K. Shaalan. 2020. A survey of semantic analysis approaches. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, pages 61–70. Springer.

Shanahan, M. 2022. Talking about large language models. *arXiv preprint arXiv:2212.03551.*

Shin, J. and J. Nam. 2021. A survey of automatic code generation from natural language. *Journal of Information Processing Systems*, 17(3):537–555.

Song, G., Y. Ye, X. Du, X. Huang, and S. Bie. 2014. Short text classification: a survey. *Journal of multimedia*, 9(5).

Srivastava, A., A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615.*

Storks, S., Q. Gao, and J. Y. Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172.*

Sun, J., S. Wang, J. Zhang, and C. Zong. 2020. Distill and replay for continual language learning. In *Proceedings of the 28th international conference on computational linguistics*, pages 3569–3579.

Tang, R., Y.-N. Chuang, and X. Hu. 2023. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205.*

Tedeschi, S., J. Bos, T. Declerck, J. Hajic, D. Hershcovich, E. H. Hovy, A. Koller, S. Krek, S. Schockaert, R. Sennrich, et al. 2023. What's the meaning of superhuman performance in today's nlu? *arXiv preprint arXiv:2305.08414.*

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971.*

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, A., Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461.*

Wang, X., J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171.*

Wang, Y., Y. Wang, J. Liu, and Z. Liu. 2020. A comprehensive survey of grammar error correction. *arXiv preprint arXiv:2005.06600.*

Wei, J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682.*

Xu, F. F., U. Alon, G. Neubig, and V. J. Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pages 1–10.

Xu, P., W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo. 2023a. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265.*

Xu, X., K. Kong, N. Liu, L. Cui, D. Wang, J. Zhang, and M. Kankanhalli. 2023b. An llm can fool itself: A prompt-based adversarial attack. *arXiv preprint arXiv:2310.13345.*

Zellers, R., A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830.*

Zhai, Y., S. Tong, X. Li, M. Cai, Q. Qu, Y. J. Lee, and Y. Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313.*

Zhang, L., S. Wang, and B. Liu. 2018. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

Zhang, R., J. Han, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, P. Gao, and Y. Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199.*

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. 2023. A survey of large language models. *arXiv e-prints*, pages arXiv–2303.

Zhong, W., R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364.*

Zhu, K., J. Wang, J. Zhou, Z. Wang, H. Chen, Y. Wang, L. Yang, W. Ye, N. Z. Gong, Y. Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528.*

Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 31-43 recibido 30-11-2023 revisado 18-01-2024 aceptado 01-02-2024

Enhancing the understanding of clinical trials with a sentence-level simplification dataset

Mejora de la comprensión de ensayos clínicos con un conjunto de datos simplificados a nivel de frase

Leonardo Campillos-Llanos,[1] Rocío Bartolomé,[2] Ana R. Terroba Reinares[3]
[1]ILLA (CSIC) [2]Fac. Filosofía y Letras (UAM) [3]Fund. Rioja Salud
leonardo.campillos@csic.es, rocio.bartolome@uam.es, arterroba@riojasalud.es

Abstract: We introduce a dataset with 1200 manually simplified sentences (144 019 tokens) from clinical trials in Spanish. A total of 1040 announcements from the European Clinical Trials Register (EudraCT) were analyzed to select sentences with ambiguities or exceeding 25 words. Simplification criteria were devised in an annotation guideline, which is released publicly along with the dataset. We obtained two versions: syntactically simplified sentences, and sentences with syntactic and lexical simplification. We report a quantitative, a qualitative and a human evaluation, in which three independent evaluators assessed the grammaticality/fluency, semantic adequacy and overall simplification. Results show that the resource is suitable for advancing research on automatic simplification of medical texts.
Keywords: Text simplification, Medical language processing, Clinical trials.

Resumen: Se presenta un conjunto de 1200 frases de ensayos clínicos en español simplificadas manualmente (144 019 tokens). Se analizaron 1040 anuncios del Registro Europeo de Ensayos Clínicos (EudraCT), seleccionando frases con ambigüedades o con más de 25 palabras. Se elaboraron criterios de simplificación recogidos en una guía distribuida públicamente con el conjunto de datos. Se obtuvieron dos versiones: oraciones simplificadas sintácticamente, y oraciones con simplificación léxica y sintáctica. Se presenta una evaluación cuantitativa, cualitativa y por tres evaluadores independientes sobre la gramaticalidad/fluidez, adecuación semántica y simplificación. Los resultados muestran que el recurso es adecuado para avanzar en la investigación en simplificación automática de textos médicos.
Palabras clave: Simplificación de textos, PLN médico, Ensayos clínicos.

1 Introduction

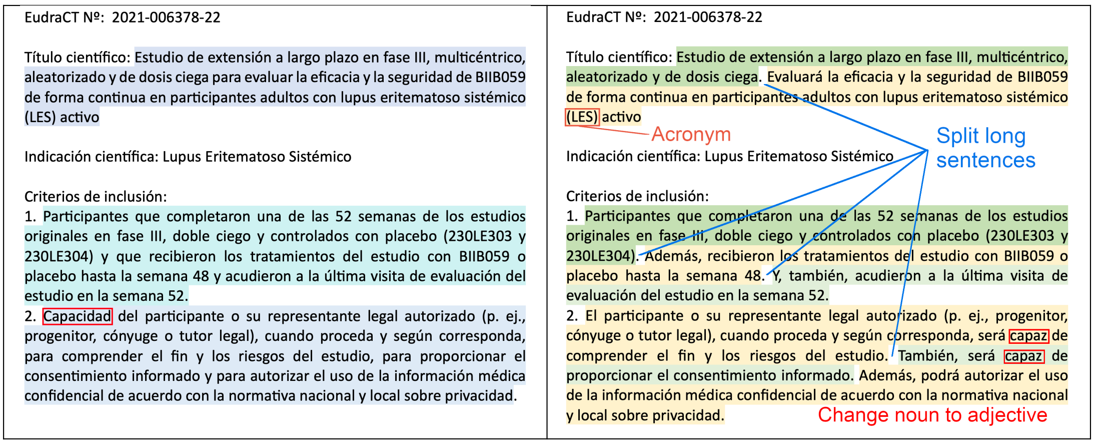
Achieving a plain language version of medical documents helps patients to enhance their understanding of health-related information and their adherence to treatment (Ondov, Attal, and Demner-Fushman, 2022). Potential participants in clinical trials might find eligibility criteria grammatically complex and rife with medical jargon (Wu et al., 2016), which hinders patients from taking part in a study. Automatic text simplification (Shardlow, 2014; Saggion, 2017), complemented with human supervision, has been shown to produce more understandable texts for patients (Lalor, Woolf, and Yu, 2019) and clinical researchers (Fang et al., 2021). Indeed, simplification also enhances (bio)medical language processing, given that such pre-processing makes it easier to parse coordinated or relative clauses (Peng et al., 2012) or complex compound phrases (Wei, Leaman, and Lu, 2014) before text mining.

To develop simplification systems for medical texts in Spanish, we created a dataset of 1200 manually simplified sentences from trial announcements. We release publicly a guideline and the resource in two versions: simplified sentences at the syntax-level, and with lexical and syntactical simplification.[1]

Figure 1 shows a sample of the original version of a trial announcement and its syntactical simplification. Long sentences in the technical version are shortened or split in the simplified version. Some nominalizations are changed to a verb or adjective form, which are easier to understand: e.g. *capacidad del*

[1]https://digital.csic.es/handle/10261/346579

ISSN 1135-5948 DOI 10.26342/2024-72-2


EudraCT Nº: 2021-006378-22

Título científico: Estudio de extensión a largo plazo en fase III, multicéntrico, aleatorizado y de dosis ciega para evaluar la eficacia y la seguridad de BIIB059 de forma continua en participantes adultos con lupus eritematoso sistémico (LES) activo

Indicación científica: Lupus Eritematoso Sistémico

Criterios de inclusión:
1. Participantes que completaron una de las 52 semanas de los estudios originales en fase III, doble ciego y controlados con placebo (230LE303 y 230LE304) y que recibieron los tratamientos del estudio con BIIB059 o placebo hasta la semana 48 y acudieron a la última visita de evaluación del estudio en la semana 52.
2. Capacidad del participante o su representante legal autorizado (p. ej., progenitor, cónyuge o tutor legal), cuando proceda y según corresponda, para comprender el fin y los riesgos del estudio, para proporcionar el consentimiento informado y para autorizar el uso de la información médica confidencial de acuerdo con la normativa nacional y local sobre privacidad.

EudraCT Nº: 2021-006378-22

Título científico: Estudio de extensión a largo plazo en fase III, multicéntrico, aleatorizado y de dosis ciega. Evaluará la eficacia y la seguridad de BIIB059 de forma continua en participantes adultos con lupus eritematoso sistémico (LES) activo

Indicación científica: Lupus Eritematoso Sistémico

Criterios de inclusión:
1. Participantes que completaron una de las 52 semanas de los estudios originales en fase III, doble ciego y controlados con placebo (230LE303 y 230LE304). Además, recibieron los tratamientos del estudio con BIIB059 o placebo hasta la semana 48. Y, también, acudieron a la última visita de evaluación del estudio en la semana 52.
2. El participante o su representante legal autorizado (p. ej., progenitor, cónyuge o tutor legal), cuando proceda y según corresponda, será capaz de comprender el fin y los riesgos del estudio. También, será capaz de proporcionar el consentimiento informado. Además, podrá autorizar el uso de la información médica confidencial de acuerdo con la normativa nacional y local sobre privacidad.

Figure 1: An unsimplified trial announcement (left) and its syntactic simplification (right).

participante... ('ability of the participant...') → *el participante será capaz...* ('the participant will be able to...'). Still, acronyms (*LES*) and technical terms (*multicéntrico*) require lexical simplification. The next sections report the background (§2), the methods (§3) and the evaluation (§4).

2 Background

Text simplification involves operations at all linguistic levels (lexis, syntax and discourse).

Methods for lexical simplification (Paetzold and Specia, 2017) generally rely on curated lexicons with technical and simplified words (Grabar and Hamon, 2016), paraphrase extraction (Elhadad and Sutaria, 2007; Deléger and Zweigenbaum, 2009), or machine learning-based approaches (Shardlow, 2013). Currently, deep learning methods are gaining ground through word-embeddings, prompt-based methods and large language models (LLMs), as explained in a recent survey (North et al., 2023). Lexical simplification has been addressed in the recent TSAR challenge (Saggion et al., 2023).

Syntactic simplification requires arranging words to achieve a word order with unambiguous references, split long sentences, change passive to active voice or rewrite nominalization structures to verb or adjective forms. Several works have used rules learned from corpora in order to apply simplification operations (Siddharthan, 2006; Peng et al., 2012; Seretan, 2012; Collados, 2013; Brouwers et al., 2014; Mukherjee et al., 2017). Most rules rely on dependency or part-of-speech tagging to derive simplification rules; for example, by parsing parallel sentences from technical and simplified texts (Szep et al., 2019). In contrast, other methods propose detecting syntactic simplification cues that do not rely on heavy syntactic analysis (Evans and Orăsan, 2019).

Discourse phenomena also require syntactic operations to simplify structures beyond the sentence and abridge long paragraphs. In addition, anaphora and co-reference might cause ambiguities to understand the content (Wilkens, Oberle, and Todirascu, 2020).

Lastly, texts may be simplified at all levels using transfer learning techniques (Menta and García-Serrano, 2022; Trienes et al., 2022; Alarcón, Martínez, and Moreno, 2023).

Evaluating simplification may be subjective (Grabar and Saggion, 2022), but standardized methods exist. However, quantitative approaches, such as readability formulae (Flesch, 1948), are not always adequate for medical texts (Zeng-Treitler et al., 2007). Moreover, metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) or SARI (Xu et al., 2016) are limited when assessing simplification, since they may correlate negatively with simplicity (Sulem, Abend, and Rappoport, 2018) or do not assess simplification operations thoroughly (Alva-Manchego, Scarton, and Specia, 2021). Human assessment of simplifications is thus beneficial.

Simplification tasks rely on lexicons or parallel (technical/simplified) corpora, which are scarce for Spanish (Segura-Bedmar and Martínez, 2017; Ferrés and Saggion, 2022; Alarcon, Moreno, and Martínez, 2023). Some were created in multilingual projects but are small (Xu, Callison-Burch, and Napoles, 2015; Martin et al., 2021; Joseph et al., 2023). We introduce a dataset to develop and test simplification tools. Table 1 shows samples.

Original	*Ensayo clínico para establecer los efectos de las dosis bajas de rtPA y los efectos de la reducción intensiva de la presión arterial en pacientes con accidente cerebrovascular isquémico agudo.* (2014-002823-86) 'Clinical trial to establish the effects of low-dose rtPA and the effects of intensive blood pressure lowering in patients with acute cerebrovascular accident'
Syntactic simplification	*Ensayo clínico para establecer los efectos de las dosis bajas de* ***rtPA*** *y los efectos de* ***reducir intensamente*** *la presión arterial. Se estudiará en pacientes con accidente cerebrovascular isquémico agudo.* 'Clinical trial to establish the effects of low-dose rtPA and the effects of lowering blood pressure intensively. This will be studied in patients with acute cerebrovascular accident'
Lexical and syntactic simplification	*Ensayo clínico para establecer los efectos de las dosis bajas de rtPA y los efectos de reducir intensamente la presión arterial.* ***rTPA es el activador recombinante del plasminógeno tisular****, un medicamento que ayuda a disolver los coágulos de sangre. Se estudiará en pacientes con accidente cerebrovascular isquémico agudo (****ictus****).* 'Clinical trial to establish the effects of low-dose rtPA and the effects of lowering blood pressure intensively. rtPA stands for recombinant tissue plasminogen activator, a medical drug that helps to dissolve blood clots. This will be studied in patients with acute cerebrovascular accident (stroke)'
Original	*Mujeres en tratamiento de TRA que reciban embriones propios o donados que presenten un desarrollo endometrial inferior a 5 mm a pesar de haber recibido un tratamiento con estrogenoterapia.* (2016-001716-38) 'Women in ART treatment that receive own or donated embryos and presenting an endometrial development less than 5 mm despite having received treatment with estrogen therapy.'
Syntactic simplification	*Mujeres en tratamiento de* ***TRA*** *que reciban embriones propios o donados. Las mujeres presentarán un desarrollo* ***endometrial*** *inferior a 5 mm a pesar de haber recibido un tratamiento con* ***estrogenoterapia****.* 'Women in ART treatment that receive own or donated embryos. These women will have an endometrial development less than 5 mm despite having received treatment with estrogen therapy.'
Lexical and syntactic simplification	*Mujeres en* ***tratamiento de reproducción asistida*** *que reciban embriones propios o donados. Las mujeres presentarán un desarrollo del* ***endometrio (capa del útero)*** *inferior a 5 mm a pesar de haber recibido una* ***terapia de estrógenos (hormonas)****.* 'Women in assisted reproductive treatment that receive own or donated embryos. These women will have a development of endometrium (the innermost layer of the uterus) less than 5 mm despite having received treatment with estrogens (hormones).'

Table 1: Samples of technical sentences and manually simplified (EudraCT id in brackets).

3 Methods

3.1 Data preparation

Three linguists analyzed trial announcements from EudraCT.[2] A set of 700 texts come from (Campillos-Llanos et al., 2021) and cover the period 2009-2020; and another set contains 340 texts (issued in the years 2020-2022). In total, we analyzed 1040 texts. However, we only used 510 trials (49.04%), because we discarded texts that were too long (above 1500 tokens), had lists with more than 10 lab values or had sentences that could not be simplified syntactically. Sentences with co-reference ambiguities, digressions or exceeding 25 words were selected (we followed a criterion supported by experts in Plain Language (da Cunha, 2022)). The criteria are detailed in §3.2 and Tables 2 and 3.

[2] `https://www.clinicaltrialsregister.eu`

`APPO`	**Appositive phrases**	
Orig	*Sujetos, **varones y mujeres**, con diagnóstico de insuficiencia renal.* ('Subjects, men and women, diagnosed with renal failure.')	(2014-001296-32)
Simp	*Sujetos con diagnóstico de insuficiencia renal.* ('Subjects diagnosed with renal failure.')	
`CONJ`	**Conjunctions (coordination and subordination)**	
Orig	*Diagnóstico por la imagen mediante fármacos radiactivos **con el objetivo de** localizar glándulas paratiroideas anómalas cuando las pruebas de imagen convencionales son negativas **y así** poder planificar de forma óptima el tratamiento quirúrgico.* ('Diagnostic imaging using radioactive pharmaceuticals to locate abnormal parathyroid glands when conventional imaging tests are negative, as a necessary condition for planning an optimal surgical treatment.')	(2019-002729-31)
Simp	*Diagnóstico por la imagen mediante fármacos radiactivos **para** localizar glándulas paratiroideas anómalas cuando las pruebas de imagen convencionales son negativas. **Así** se podrá planificar de forma óptima el tratamiento quirúrgico.* ('Diagnostic imaging using radioactive pharmaceuticals to locate abnormal parathyroid glands when conventional imaging tests are negative. Surgical treatment can then be optimally planned.')	
`COREF`	**Co-reference and anaphora**	
Orig	*Ensayo clínico para la identificación de biomarcadores basados en técnicas ómicas (..), y su variabilidad inter e intraindividual **que** permitan la mejora en la individualización del tratamiento.* ('Clinical trial for the identification of biomarkers based on omics techniques (..), and their inter and intra-individual variability that allow the improvement in the individualization of treatment.')	(2019-002795-13)
Simp	*Ensayo clínico para identificar biomarcadores basados en técnicas ómicas (..), y su variabilidad inter e intraindividual. **Estas técnicas** permitirían la mejora en la individualización del tratamiento.* ('Clinical trial to identify biomarkers based on omics techniques (..), and their inter- and intra-individual variability. These techniques would allow the improvement in the individualization of treatment.')	
`LEN`	**Long sentences**	
Orig	*Las pacientes fértiles deberán obtener resultado negativo en una prueba de embarazo en orina en las 24 horas previas a la primera dosis del fármaco del estudio.* ('Female subjects of childbearing potential must have a negative urine pregnancy test within 24 hours prior to the first dose of study drug.')	(2019-001565-33)
Simp	*Las pacientes fértiles deberán obtener resultado negativo en una prueba de embarazo en orina. Se hará en las 24 horas previas a la primera dosis del fármaco del estudio.* ('Female subjects of childbearing potential must have a negative urine pregnancy test. This will be performed within 24 hours prior to the first dose of study drug.')	
`NEG`	**Negation**	
Orig	***No más de 1 año** antes de la fecha de inclusión.* ('No more than 1 year prior to enrollment.')	(2015-003759-23)
Simp	***Un año o menos** antes de la fecha de inclusión.* ('One year or less prior to enrollment.')	

Table 2: Syntactic simplification aspects according to linguistic criteria.

NOM	**Change nouns/adjectives to verb form**	
Orig:	*Paracetamol en el* ***tratamiento*** *del dolor.* ('Paracetamol in the treatment of pain.')	(2015-004482-88)
Simp:	*Paracetamol para* ***tratar*** *el dolor.* ('Paracetamol to treat pain.')	
PAS	**Passive to active voice**	
Orig:	*4 semanas previas a dosificación, o más si* ***es requerido*** *por las regulaciones locales.* ('4 weeks before dosage, or more if it is required by local regulations')	(2016-001227-31)
Simp:	*4 semanas previas a dosificación, o más si las regulaciones locales lo* ***requieren****.* ('4 weeks before dosage, or more if local regulations require it.')	
REDUN	**Redundancies**	
Orig	*Se debe consultar al monitor médico antes de que el participante del* ***estudio*** *se incorpore al* ***estudio*** *AS0014.* ('The medical monitor must be consulted prior to the study participant's entry into the AS0014 study.')	(2019-004163-47)
Simp	*Se debe consultar al monitor médico antes de que el participante se incorpore al* ***estudio*** *AS0014.* ('The medical monitor must be consulted before the participant enters into the AS0014 study.')	
OVERS	**Oversimplification**	
Orig	*En participantes sintomáticos, uno de los criterios para el diagnóstico de posible demencia frontotemporal de variante conductual o de subtipo semántico o de afasia progresiva primaria.* ('In symptomatic patients, one of the criteria for the diagnosis of probable behavioral variant FTD or FTD-semantic subtype or FTD-Progressive Non-fluent Aphasia.')	(2019-004066-18)
Simp	*En participantes sintomáticos,* ***que tengan*** *uno de los criterios para el diagnóstico de posible demencia frontotemporal de variante conductual.* ***También, que tengan*** *posible demencia frontotemporal de subtipo semántico o de afasia progresiva primaria.* ('In symptomatic patients, participants who have one of the criteria for the diagnosis of possible behavioral variant of frontotemporal dementia. Also, participants who have possible frontotemporal dementia of semantic subtype or primary progressive aphasia.')	
OTHER	**Other:** This label gathers aspects related to style, punctuation or grammar that enhance the clarity of the sentence or avoid ambiguities; these operations include fixing number or gender disagreement, preposition errors or unnatural word order.	
Orig	*Aborto recurrente, preeclampsia previa o enfermedades hematológicas. Uso de* ***fármacos vasoactivos****: Fundamentalmente* ***relacionadas*** *con la hipertensión.* ('Recurrent miscarriage, previous preeclampsia or hematologic diseases. Use of vasoactive drugs: Fundamentally related to hypertension.')	(2017-001878-42)
Simp	*Aborto recurrente, preeclampsia previa o enfermedades hematológicas. Uso de fármacos vasoactivos fundamentalmente relacionados con la hipertensión.* ('Recurrent miscarriage, previous preeclampsia or hematologic diseases. Use of vasoactive drugs mainly related to hypertension.')	

Table 2: Syntactic simplification aspects according to linguistic criteria (cont.).

3.2 Simplification criteria

We followed the works by experts in plain language (da Cunha, 2022), the recommendations of the International Plain Language Federation,[3] the guideline prepared by the European Commission (European-Commission, 2016) and lexical simplification analyses (Koptient, Cardon, and Grabar, 2019; Carbajo and Moreno-Sandoval, 2023). We also applied the criteria defined in former work (Campillos-Llanos et al., 2022).

We provide two versions: syntactically simplified sentences, and sentences with syntactic and lexical simplifications. The version without lexical simplification is intended for research on syntactic simplification (e.g. development of a dedicated tool). The fully simplified one is provided for end-to-end systems that simplify sentences at all levels. A guideline gathers the simplification criteria.[4] Tables 2 and 3 show all simplification aspects.

3.3 Analysis and evaluation

To understand the distribution of topics across sentences, we used Medical Subject Heading (MeSH) Tree Entry Terms from the corresponding source text. Each EudraCT trial announcement has a MeSH descriptor (section E.1.1.2) of the therapeutic area. Nonetheless, these are not always accurate, and our topic distribution is only illustrative. We also counted the most frequent medical concepts in the sentences. Although the distributed dataset is not normalized to Concept Unique Identifiers from the Unified Medical Language System (Bodenreider, 2004), we used a lexicon (Campillos-Llanos, 2023) for the normalization used in this analysis.

To measure the quality of our simplifications, we conducted quantitative and qualitative measurements. First, we compared the word count, the number of syllables per sentence, the count of polysyllable words (with at least 3 syllables) and of monosyllable words in original and simplified sentences. We used `Textstat` (Bansal and Aggarwal, 2021). Simplified sentences should be shorter, have less syllables or less polysyllable words. We also compared the dependency tree height. This is a measure of structural complexity, given that more complex sentences have deeper syntactic dependency trees, as other teams showed (Alva-Manchego et al., 2020; Martin et al., 2020). The dependency tree depth should be shorter in simplified sentences. We computed this value with the Spacy `es_core_news_sm` model (vs. 3.3.3). For example, the following unsimplified sentence (with an apposition) has a token count of 7 and a dependency tree height value of 3: *Subjects, men and women, with renal insufficiency.* In contrast, the simplified version (without the apposition) has less tokens (4) and a shallower dependency tree height (2): *Subjects with renal insufficiency.* Figure 2 shows the dependency parsing of both sentences (obtained with Spacy).

To compare the lexical diversity of each simplification version, we computed the type-token ratio (TTR), a measure that has been used to describe other corpora (Trienes et al., 2022). The TTR is the proportion between unique tokens (*types*) and all tokens in a corpus. The higher the TTR value, the more lexically diverse a text is, and presumably more complex. We used a Python script.[5]

As a proxy for readability, we computed the average Inflesz score for each version (original, syntactic simplification and lexical and syntactic simplification). This is a perspicuity-based measure to estimate how clear and comprehensible a text is, according to the count of words, syllables and sentences. The Inflesz value was validated in Spanish health texts (Barrio-Cantalejo et al., 2008). The higher the score, the more readable the text is. We used a Python implementation.[6]

We did not use BLEU nor SARI since we did not compare the output of any simplification method with the human simplifications.

For a qualitative evaluation, three subjects (one linguist and two documentalists who were not involved in the simplification) assessed 100 random simplified sentences (50 with syntax simplification, and 50 with both the syntax and lexical simplification). They evaluated grammaticality and fluency (G/F), semantic coherence and meaning adequacy (M), and overall simplification (S), in line with previous work (Saggion et al., 2015; Koptient and Grabar, 2020). A 5-point Likert scale questionnaire was distributed (5 was the highest score, and 1 the lowest). We modified instructions originally prepared by other teams (Yamaguchi et al., 2023) to fit the Spanish language.

[3] `https://www.iplfederation.org/`

[4] `https://digital.csic.es/handle/10261/346579`

[5] Available at: `https://acortar.link/N49259`

[6] Available at: `https://acortar.link/9i8yF0`

`ABBR`	**Expanding abbreviations/acronyms**	
Orig:	*Tratamiento para el* ***MDE****.* ('Treatment for MDE.')	(2019-002704-41)
Simp:	*Tratamiento para el* ***episodio depresivo mayor****.* ('Treatment for mayor depressive episode.')	
`ADD-LEX`	**Adding a lexeme**	
Orig:	*Tolerabilidad de macitentan.* ('Tolerability of macitentan.')	(2013-003822-96)
Simp:	*Tolerabilidad del* ***medicamento*** *macitentan* ('Tolerability of macitentan medical drug.')	
`DEL-LEX`	**Deleting a lexeme**	
Orig	*Elvitegravir* ***(EVG)*** *administrado junto a darunavir.* ('Elvitegravir (EVG) administered with darunavir.')	(2013-001476-37)
Simp	*Elvitegravir administrado junto a darunavir.* ('Elvitegravir administered with darunavir.')	
`HYP`	**Replacement with a hypernym**	
Orig	*Ensayo clínico, simple ciego, aleatorizado, controlado y prospectivo.* ('Single blind, randomized, controlled prospective clinical trial.')	(2012-005571-14)
Simp	***Ensayo clínico de investigación.*** ('Clinical research trial.')	
`PAR`	**Paraphrase or definition**	
Orig	*Tratamiento con amikacina* ***intravenosa****.* ('Treatment with intravenous amikacine.')	(2014-001296-32)
Simp	*Tratamiento con amikacina* ***administrada en vena****.* ('Treatment with amikacine administered into the vein.')	
`SYN`	**Simpler synonym**	
Orig	***Profilaxis*** *habitual.* ('Usual prophylaxis.')	(2019-002233-11)
Simp	***Prevención*** *habitual.* ('Usual prevention.')	
`TRANS`	**Translation**	
Orig	*Test de embarazo de la visita de* ***screening****.* ('Pregnancy test at the screening visit.')	(2020-001901-22)
Simp	*Test de embarazo de la visita de* ***selección****.*	

Table 3: Lexical simplification aspects according to linguistic criteria.

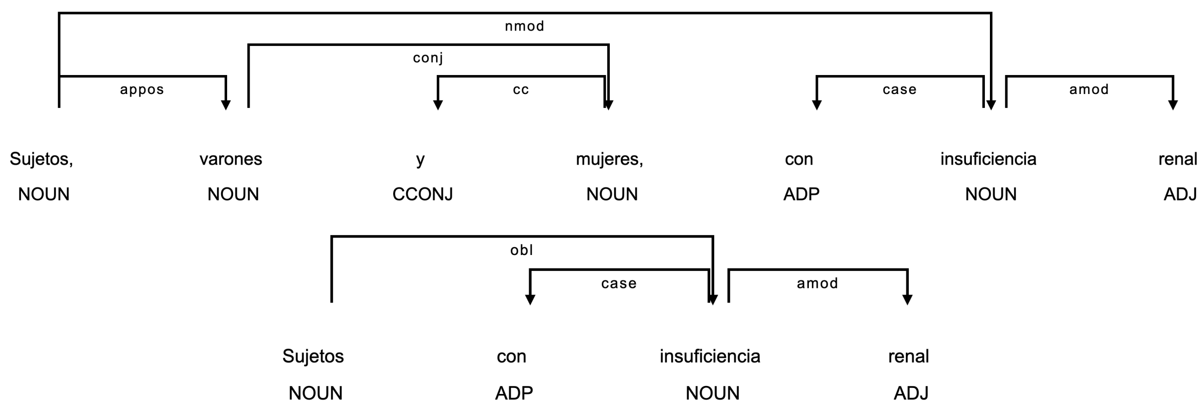


Figure 2: Dependency parsing of an unsimplified sentence with an apposition (*appos*) above and the simplified sentence (without the apposition) below; *ADP*: 'adposition' (∼preposition).

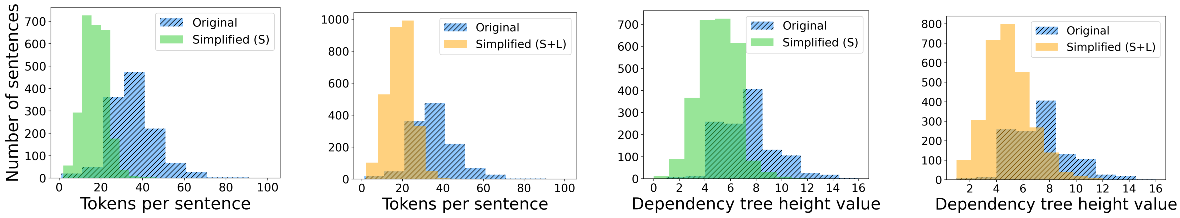


Figure 3: Tokens per sentence and dependency tree height values of original sentences, sentences with syntactic simplification (S) and with syntactic and lexical simplification (S+L).

	Original	**S**	**S+L**
Tokens (tk)	43 229	45 013	55 777
Types	5625	5669	5764
TTR	0.13	0.10	0.12
Avg tk/st	35.66 (±10.51)	17.31 (±3.91)	19.16 (±3.93)
Avg st	1.02 (±0.16)	2.21 (±0.62)	2.47 (±0.85)
Avg syl/st	66.55 (±19.45)	31.50 (±10.57)	34.48 (±11.97)
Avg mon/st	19.40 (±6.64)	9.03 (±3.87)	9.89 (±4.13)
Avg pol/st	8.63 (±2.65)	4.34 (±2.12)	4.61 (±2.37)

Table 4: Counts; *S*: syntactic simplification; *S+L*: syntactic and lexical simplification; *Avg*: average; *TTR*: type-token ratio; *st*: sentence; *syl*: syllables; *mon*: monosyllable words; *pol*: polysyllable.

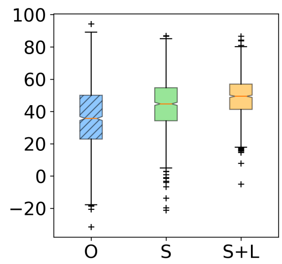


Figure 4: Inflesz scores.

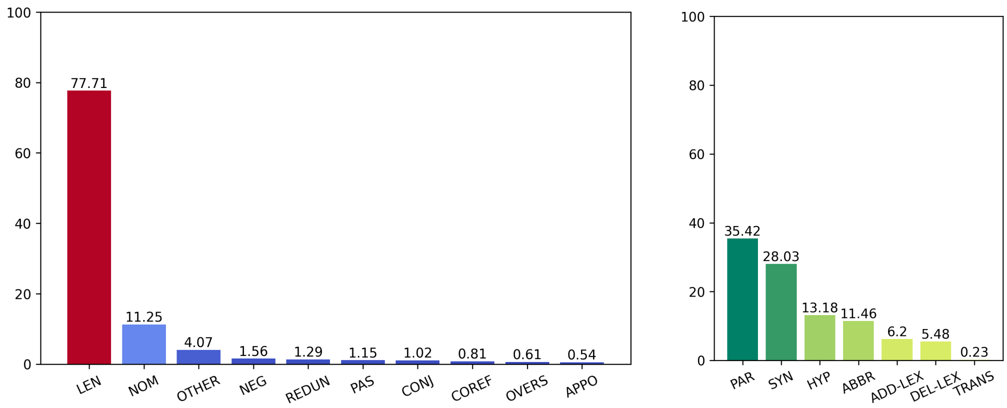


Figure 5: Distribution (%) of syntactic (left) and lexical simplification operations (right).

4 Results

4.1 Descriptive statistics

Table 4 shows descriptive statistics. The syntactically simplified version has shorter sentences and contain less syllables and polysyllable words. However, monosyllable words are more abundant in the original version. The number of simplified sentences tends to be slightly superior to the original version; indeed, many simplifications involved splitting long sentences. The fully-simplified version contains marginally longer sentences, more syllables and polysyllable words compared to the version with only syntactic simplification. Similarly, the TTR scores were lower in the syntactically simplified version. The original and the fully-simplified versions were more lexically diverse, possibly because they have more jargon or paraphrases, respectively.

With regard to readability (Figure 4), the average Inflesz score of the original version was of 35.69 (±19.72), which is interpreted as *Very difficult*. The syntactically simplified version has a higher score (44.19±15.42); and sentences with both syntactic and lexical simplification have a higher score (48.99±11.75). Scores of both simplified versions are considered *Somewhat difficult* in Inflesz.

Regarding the dependency tree height, the average value was of 7.13 in the original sentences; 4.80 in the syntactically-simplified ones; and 5.06 in the sentences both syntactically and lexically simplified. Figure 3 shows the distribution of word count and dependency tree height values across versions. Statistical tests of these values, count of syllables, monosyllable and polysyllable words across versions showed statistically significant differences (Kruskall-Wallis, p<0.0001).

Overall, lexical aspects needed more simplification. Figure 5 shows the distribution of simplification operations. A total of 1476 syntactic aspects were simplified (an average of 1.23 operations per sentence). Shortening sentence length and changing nominal/adjective structures to verbs were the most frequent operations in our data. In turn, 2208 lexical aspects were simplified (an average of 1.84). Altogether, semantic-related lexical changes (paraphrasing, synonym and hypernym replacement) are estimated to represent up to 76.63% of lexical operations. Abbreviations/acronyms account for a 11.46%.

4.2 Health topics and concepts

Table 5 shows the 15 most frequent UMLS Concept Unique Identifiers (CUIs). Most refer to research tasks (*clinical trials, evaluation, randomization*), participants (*patients, study subjects*) and general entities about conditions or procedures (*disease, pharmaceutical preparations (prep.), medicament*).

Freq	CUI and preferred term
2604	C0008976; Clinical Trials
2122	C0008972; Clinical Research
1875	C2603343; Study
1719	C0030705; Patients
1372	C0087111; Therapeutic procedure
857	C0220825; Evaluation
529	C0681850; Study Subject
490	C0034656; Randomization
438	C0008976; Clinical Trials
412	C0012634; Disease
353	C0013227; Pharmaceutical prep.
350	C0221423; Illness
348	C1510438; Assay
337	C0456386; Medicament
337	C0304228; Proprietary drug

Table 5: The most frequent CUIs.

Figure 6 plots the topic distribution of Medical Subject Headings (MeSH) in the EudraCT source texts (15 most frequent topics). Most sentences come from clinical trial announcements about cancer (19.69%), virus diseases (11.42%), nervous system diseases (9.25%) and cardiovascular diseases (7.48%).

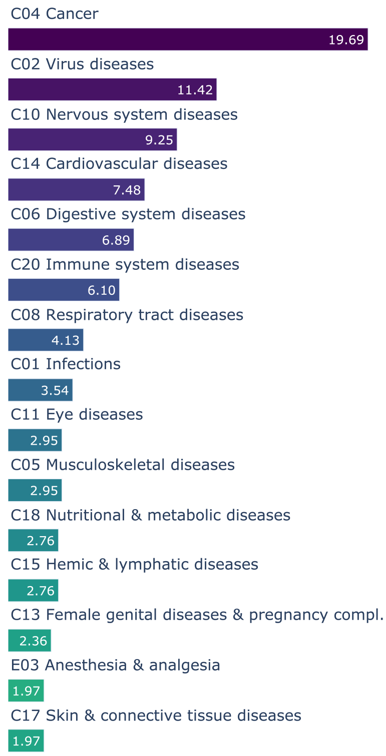


Figure 6: The 15 most frequent MeSH topics (%) in the EudraCT source texts.

	G/F	M	Sim	Avg
S	4.9	4.9	3.6	4.5
S+L	4.8	4.9	4.3	4.7

Table 6: Human evaluation; *G/F*: grammaticaly/fluency; *M*: meaning; *Sim*: simplification; *S*: syntactically simplified; *S+L*: syntactically and lexically simplified; *Avg*: average.

4.3 Human evaluation

Table 6 includes the evaluation results; we also include the average of the three aspects as in (Maddela, Alva-Manchego, and Xu, 2021). The average simplification scores (*Sim*) were slightly lower in the version with only syntactical simplification. Grammaticality and fluency aspects (*G/F*) were moderately similar. Still, some sentences from the version with both lexical and syntactic simplification were penalized due to many explanations in brackets that decreased the perceived readability. However, with regard

O: *Se estudiará* ***en el tratamiento de*** *pacientes* ((PRO) (V)) (PP ((PREP) (NP ((DET)(N) (PP ((PREP) (NP (N))))))	(NOM)
S: *Se estudiará* ***para tratar*** *pacientes* ((PRO) (V)) (PP (PREP) ((V) (NP (NOUN)))	
O: *Sujetos,* ***varones y mujeres,*** *con insuficiencia renal* (NP ((N) (PUNCT) (NP (N) (CCONJ) (N)) (PUNCT) (PP (PREP) (NP (N) (AP (ADJ))))))	(APPO)
S: *Sujetos con insuficiencia renal* (NP ((N) (PP (PREP) (NP (N) (AP (ADJ))))))	

Table 7: Samples of rules to change original (O) to simplified sentences (S). ADJ: 'adjective'; AP: 'adjective phrase'; DET: 'determiner'; N: 'noun'; NP: 'noun phrase'; PRO: 'pronoun'; PP: 'prepositional phrase'; $PREP$: 'preposition'; $PUNCT$: 'punctuation'; V: 'verb'.

to simplification, the fully simplified version received higher scores. This implies that both syntactic and lexical aspects achieved the best overall simplification, although we need to improve sentence fluency.

5 Discussion

Readers can not always understand medical documents due to long sentences, terminology and opaque acronyms. Simplifying medical texts needs to address syntactic and lexical aspects. However, any simplification task poses the challenge of guaranteeing to transmit the meaning of the text with precision.

We present a manually-simplified dataset for automatic simplification of medical texts. Our quantitative evaluation showed that sentence length, average tokens, syllables and polysyllable words per sentence, and dependency tree height values were lower in simplified sentences—i.e., these are less complex. Inflesz readability scores showed that simplified sentences are less difficult; still, according to this scale, they are *Somewhat difficult*. This is in line with our human evaluation, in which the overall simplification was rated in a 5-point Likert scale with lower scores (compared to fluency or semantic adequacy). All in all, there is still room for improvement, but the version with syntactic and lexical simplifications was rated better on average.

Our work has several limitations. First, the dataset size is small, which makes it difficult to train data-intensive approaches. More sentences need to be simplified by humans, which is a labor-intensive task. Second, the human evaluation could be subjective and is not strong (only 3 subjects assessed 100 simplified sentences, due to time constraints). Third, we did not test any syntactic simplification system. Some tools are only available for English (Mukherjee et al., 2017; Scarton et al., 2017; Chatterjee and Agarwal, 2021). Other tools for Spanish are not openly accessible (Ferrés et al., 2016). Creating (or re-adapting) a system for the Spanish language is out the scope of the present work. Lastly, although we described linguistic simplification aspects, we did not annotate abstract operations (e.g. `delete`, `add`, `move` or `replace`), as in other works (Bott and Saggion, 2011; Cardon et al., 2022).

On the whole, this is one of the few available resources for medical text simplification in Spanish. The dataset can be used to derive sentence-level simplification rules. Part-of-speech tagging the original and simplified sentences allows linguists to extract rules across registers, as other teams did (Seretan, 2012; Szep et al., 2019). Table 7 illustrates some samples of simplification rules.

6 Conclusion

We presented a dataset of 1200 sentences from clinical trial announcements in Spanish. Three experts simplified them manually according to criteria recorded in a guideline, which is shared publicly along with the dataset. We distribute a syntactically simplified version, and another with lexical and syntactic simplification. We reported descriptive statistics, an analysis of health topics and concepts, and a quantitative evaluation. A human evaluation showed that the simplified sentences are adequate, and the fully simplified version was assessed better. The main limitations are the small dataset size and the limited human evaluation.

Acknowledgements

This work was conducted in project CLARA-MED (PID2020-116001RA-C33) funded by MCIN/AEI/10.13039/501100011033/ (call: "Proyectos I+D+i Retos Investigación").

References

Alarcón, R., P. Martínez, and L. Moreno. 2023. Tuning bart models to simplify spanish health-related content. *Procesamiento del Lenguaje Natural*, 70:111–122.

Alarcon, R., L. Moreno, and P. Martínez. 2023. EASIER corpus: A lexical simplification resource for people with cognitive impairments. *Plos one*, 18(4):e0283622.

Alva-Manchego, F., L. Martin, A. Bordes, C. Scarton, B. Sagot, and L. Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proc. of the 58th ACL*, page 4668–4679.

Alva-Manchego, F., C. Scarton, and L. Specia. 2021. The (un) suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Bansal, S. and C. Aggarwal. 2021. Textstat. `https://pypi.org/project/textstat/`.

Barrio-Cantalejo, I. M., P. Simón-Lorda, M. Melguizo, I. Escalona, M. I. Marijuán, and P. Hernando. 2008. Validación de la Escala INFLESZ para evaluar la legibilidad de los textos dirigidos a pacientes. 31(2):135–152.

Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Bott, S. M. and H. Saggion. 2011. Spanish text simplification: An exploratory study. *Procesamiento del Lenguaje Natural*, 47:87–95.

Brouwers, L., D. Bernhard, A.-L. Ligozat, and T. François. 2014. Syntactic sentence simplification for French. In *Proc. of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 47–56.

Campillos-Llanos, L., A. R. Terroba Reinares, S. Zakhir Puig, A. Valverde-Mateos, and A. Capllonch-Carrión. 2022. Building a comparable corpus and a benchmark for Spanish medical text simplification. *Procesamiento del lenguaje natural*, pages 189–196.

Campillos-Llanos, L. 2023. MedLexSp–a medical lexicon for Spanish medical natural language processing. *Journal of Biomedical Semantics*, 14(1):1–23.

Campillos-Llanos, L., A. Valverde-Mateos, A. Capllonch-Carrión, and A. Moreno-Sandoval. 2021. A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC Med Inform Decis Mak*, 21(1):1–19.

Carbajo, B. and A. Moreno-Sandoval. 2023. Financial concepts extraction and lexical simplification in spanish. *(Under review)*.

Cardon, R., A. Bibal, R. Wilkens, D. Alfter, M. Norré, A. Müller, W. Patrick, and T. François. 2022. Linguistic corpus annotation for automatic text simplification evaluation. In *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1842–1866.

Chatterjee, N. and R. Agarwal. 2021. DEPSYM: A Lightweight Syntactic Text Simplification Approach using Dependency Trees. In *CTTS@ SEPLN*, pages 42–56.

Collados, J. C. 2013. Splitting complex sentences for natural language processing applications: Building a simplified Spanish corpus. *Procedia-Social and Behavioral Sciences*, 95:464–472.

da Cunha, I. 2022. Un redactor asistido para adaptar textos administrativos a lenguaje claro. *Procesamiento del Lenguaje Natural*, 69:39–49.

Deléger, L. and P. Zweigenbaum. 2009. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proc. of the 2nd Workshop on Building and Using Comparable Corpora*, pages 2–10.

Elhadad, N. and K. Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *Biological, translational, and clinical language processing*, pages 49–56.

European-Commission. 2016. *Cómo escribir con claridad*. Brussels: Directorate-General for Translation, Publications Office.

Evans, R. and C. Orăsan. 2019. Identifying signs of syntactic complexity for rule-based sentence simplification. *Natural Language Engineering*, 25(1):69–119.

Fang, Y., J. H. Kim, B. R. S. Idnay, R. A. Garcia, C. E. Castillo, Y. Sun, H. Liu, C. Liu, C. Yuan, and C. Weng. 2021. Participatory design of a clinical trial eligibility criteria simplification method. In *Medical Informatics Europe*, pages 984–988.

Ferrés, D., M. Marimon, H. Saggion, and A. AbuRa'ed. 2016. YATS: yet another text simplifier. In *Proc. of the 21st Int. Conf. on Applications of Natural Language to Information Systems, NLDB 2016*, pages 335–342. Springer.

Ferrés, D. and H. Saggion. 2022. ALEXSIS: a dataset for lexical simplification in Spanish. In *Proc. of LREC 2022*, pages 3582–94, Marseille, France.

Flesch, R. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Grabar, N. and T. Hamon. 2016. A large rated lexicon with French medical words. In *Proc. of LREC 2016*, pages 2643–2648, Portorož, Slovenia.

Grabar, N. and H. Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence TALN*, pages 453–463.

Joseph, S., K. Kazanas, K. Reina, V. J. Ramanathan, W. Xu, B. C. Wallace, and J. J. Li. 2023. Multilingual simplification of medical texts. *arXiv preprint arXiv:2305.12532*.

Koptient, A., R. Cardon, and N. Grabar. 2019. Simplification-induced transformations: typology and some characteristics. In *BioNLP 2019*, page 309–318.

Koptient, A. and N. Grabar. 2020. Fine-grained text simplification in French: steps towards a better grammaticality. In P. Bath, P. Jokela, and L. Sbaffi, editors, *Proc. of Int. Symp. on Health Information Management Research*.

Lalor, J. P., B. Woolf, and H. Yu. 2019. Improving electronic health record note comprehension with NoteAid: randomized trial of electronic health record note comprehension interventions with crowdsourced workers. *Journal of medical Internet research*, 21(1):e10793.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proc. of Workshop on Text Summarization of ACL*, pages 74–81, Barcelona, Spain.

Maddela, M., F. Alva-Manchego, and W. Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proc. of NAACL*, pages 3536–3553.

Martin, L., A. Fan, É. de la Clergerie, A. Bordes, and B. Sagot. 2021. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Martin, L., B. Sagot, E. de la Clergerie, and A. Bordes. 2020. Controllable sentence simplification. In *Proc. of LREC 2020*, pages 4689–4698, Marseille, France.

Menta, A. and A. García-Serrano. 2022. Controllable sentence simplification using transfer learning. *Proc. of the Working Notes of CLEF*.

Mukherjee, P., G. Leroy, D. Kauchak, S. Rajanarayanan, D. Y. R. Diaz, N. P. Yuan, T. G. Pritchard, and S. Colina. 2017. NegAIT: A new parser for medical text simplification using morphological, sentential and double negation. *Journal of biomedical informatics*, 69:55–62.

North, K., T. Ranasinghe, M. Shardlow, and M. Zampieri. 2023. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*.

Ondov, B., K. Attal, and D. Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988.

Paetzold, G. H. and L. Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the ACL*, pages 311–318.

Peng, Y., C. O. Tudor, M. Torii, C. H. Wu, and K. Vijay-Shanker. 2012. iSimp: A sentence simplification system for biomedical text. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1–6. IEEE.

Saggion, H. 2017. *Automatic text simplification*, volume 32. Synthesis Lectures on Human Language Technologies, Springer.

Saggion, H., S. Štajner, S. Bott, S. Mille, L. Rello, and B. Drndarevic. 2015. Making it Simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Saggion, H., S. Štajner, D. Ferrés, K. C. Sheang, M. Shardlow, K. North, and M. Zampieri. 2023. Findings of the TSAR-2022 shared task on multilingual lexical simplification. *arXiv preprint arXiv:2302.02888*.

Scarton, C., A. P. Aprosio, S. Tonelli, T. M. Wanton, and L. Specia. 2017. MUSST: A multilingual syntactic simplification tool. In *Proc. of the IJCNLP 2017, System Demonstrations*, pages 25–28.

Segura-Bedmar, I. and P. Martínez. 2017. Simplifying drug package leaflets written in Spanish by using word embedding. *Journal of biomedical semantics*, 8(1):1–9.

Seretan, V. 2012. Acquisition of Syntactic Simplification Rules for French. In *Proc. of LREC*, pages 4019–4026.

Shardlow, M. 2013. A comparison of techniques to automatically identify complex words. In *Proc. of the 51st annual meeting of the Association for Computational Linguistics*, pages 103–109.

Shardlow, M. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Siddharthan, A. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:77–109.

Sulem, E., O. Abend, and A. Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proc. of the 2018 EMNLP Conference*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Szep, A., M. Szep, G. Leroy, D. Kauchak, N. Kloehn, D. Revere, and M. Just. 2019. Algorithmic generation of grammar simplification rules using large corpora. *AMIA Summits on Translational Science Proceedings*, 2019:72–81.

Trienes, J., J. Schlötterer, H.-U. Schildhaus, and C. Seifert. 2022. Patient-friendly clinical notes: towards a new text simplification dataset. In *Proc. of the TSAR-2022 Workshop*, pages 19–27.

Wei, C.-H., R. Leaman, and Z. Lu. 2014. Simconcept: A hybrid approach for simplifying composite named entities in biomedicine. In *Proc. of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 138–146.

Wilkens, R., B. Oberle, and A. Todirascu. 2020. Coreference-based text simplification. In *Proc. of the 1st READI Workshop*, pages 93–100.

Wu, D. T., D. A. Hanauer, Q. Mei, P. M. Clark, L. C. An, J. Proulx, Q. T. Zeng, V. V. Vydiswaran, K. Collins-Thompson, and K. Zheng. 2016. Assessing the readability of clinicaltrials.gov. *Journal of the American Medical Informatics Association*, 23(2):269–275.

Xu, W., C. Callison-Burch, and C. Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Xu, W., C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yamaguchi, D., R. Miyata, S. Shimada, and S. Sato. 2023. Gauging the gap between human and machine text simplification through analytical evaluation of simplification strategies and errors. In *Findings of EACL 2023*, pages 359–375.

Zeng-Treitler, Q., H. Kim, S. Goryachev, A. Keselman, L. Slaughter, and C.-A. Smith. 2007. Text characteristics of clinical reports and their implications for the readability of personal health records. *Studies in health technology and informatics*, 129(2):1117.

Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 45-55 recibido 13-12-2023 revisado 15-01-2024 aceptado 29-01-2024

An Empirical Study on the Number of Items in Human Evaluation of Automatically Generated Texts

Estudio Empírico sobre el Número de Elementos en la Evaluación Humana de Textos Generados Automáticamente

Javier González-Corbelle,[1] Jose M. Alonso-Moral,[1] Rosa M. Crujeiras,[2] Alberto Bugarín-Diz[1]
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),
Universidade de Santiago de Compostela, Spain
[2]Centro de Investigación e Tecnoloxía Matemática de Galicia (CITMAga),
Universidade de Santiago de Compostela, Spain
{j.gonzalez.corbelle, josemaria.alonso.moral, rosa.crujeiras, alberto.bugarin.diz}@usc.es

Abstract: Human evaluation of neural models in Natural Language Generation (NLG) requires a careful experimental design in terms of the number of evaluators, number of items to assess, number of quality criteria, among other factors, for the sake of reproducibility as well as for ensuring that significant conclusions are drawn. Although there are some generic recommendations on how to proceed, there is not an established or accepted evaluation protocol admitted worldwide yet. In this paper, we address empirically the impact of the number of items to assess in the context of human evaluation of NLG systems. We first apply resampling methods to simulate the evaluation of different sets of items by each evaluator. Then, we compare the results obtained by evaluating only a limited set of items with those obtained by evaluating all outputs of the system for a given test set. Empirical findings validate the research hypothesis: well-known resampling statistical methods can contribute to getting significant results even with a small number of items to be evaluated by each evaluator.
Keywords: Natural Language Generation, Human Evaluation, Resampling Methods.

Resumen: La evaluación humana de modelos neuronales en Generación de Lenguaje Natural (GLN) requiere un diseño experimental cuidadoso de elementos como, por ejemplo, número de evaluadores, número de ítems a evaluar, número de criterios de calidad, entre otros, para así garantizar la reproducibilidad de experimentos, así como para asegurar que las conclusiones extraídas son significativas. Aunque existen algunas recomendaciones genéricas sobre cómo proceder, no existe un protocolo de evaluación consensuado, general y aceptado. En este artículo prestamos atención a cómo influye el número de elementos a evaluar en la evaluación humana de los sistemas de GLN. Aplicamos distintos métodos de remuestreo para simular la evaluación de distintos conjuntos de ítems por parte de cada evaluador. A continuación, comparamos los resultados obtenidos evaluando sólo un conjunto limitado de ítems con los obtenidos evaluando todas las salidas del sistema para el conjunto completo de casos de prueba. Las conclusiones derivadas del estudio empírico corroboran la hipótesis de investigación de partida: el uso de técnicas de remuestreo ayuda a obtener resultados de evaluación significativos incluso con un número pequeño de ítems a evaluar por cada evaluador.
Palabras clave: Generación de Lenguaje Natural, Evaluación Humana, Remuestreo.

1 Introduction

There is debate about the use of automatic metrics versus human judgement when evaluating the output of Natural Language Generation (NLG) systems. Reiter (2018) stated that commonly used automatic metrics, such as ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), or BLEU (Papineni et al., 2002), do not correlate well with human judgements for the evaluation of NLG systems. This is mainly because the most popular metrics are based on checking the n-gram overlap of the generated sentence with a limited set of reference texts that are considered correct, but do not cover all the possible text variations that NLG systems may

ISSN 1135-5948 DOI 10.26342/2024-72-3

produce (e.g., paraphrases, synonyms, or alternate realizations). Accordingly, other metrics have emerged, such us embeddings-based metric to measure similarity between reference and candidate texts like BERTScore Zhang et al. (2020), or pre-trained metrics, i.e., neural models trained to learn how to automatically do an evaluation task, like BLEURT (Sellam, Das, and Parikh, 2020) or NUBIA (Kane et al., 2020). More recently, there are studies regarding the application of ChatGPT for assessing generation tasks (Wang et al., 2023). However, despite these efforts to produce more and more data-driven automatic metrics, which are inspired from the machine learning community, the lack of correlation with human evaluation persists (Moramarco et al., 2022).

On the other hand, Van der Lee et al. (2021) recommended some best practices for human evaluation, and the NLG research community is doing efforts to set the basis for reproducible human evaluation (Belz et al., 2023; Belz, 2022). But, in spite of this, there is still a lack of formal protocol for carrying out NLG human evaluation. Furthermore, conducting human evaluation properly is not straightforward, since there are multiple factors that must be considered, being among them the textual properties to be assessed, the evaluation criteria that human evaluators must follow, the number of human evaluators, the number of items to evaluate, the number of questions per item, the statistical tests, tools for data analysis, etc.

In this paper we focus on validating the following research hypothesis: "well-known resampling statistical methods can contribute to getting significant results even with a small number of items to be evaluated by each evaluator". Thus, we aim to prove empirically the influence of the number of items presented to an evaluator in the context of human NLG evaluation. Starting from a set of texts generated by an NLG system (i.e., a set of items to be evaluated), we research on the minimal number of texts to be assessed for ensuring that the evaluation results obtained are significant.

More precisely, we apply two resampling methods to simulate multiple evaluations, thus exploring the effect of different number of items per evaluation. As far as we know, this is the first empirical study regarding the impact of the number of items to assess in NLG human evaluation. Notice that the concept of "item" may vary depending on the context. In the context of NLG evaluation, some researchers may understand as "item" each criteria used to manually evaluate the text (e.g., coherence, quality, etc.), but in this paper item refers to each text to be evaluated.

The rest of the manuscript is organized as follows. Section 2 introduces some preliminary concepts. Section 3 presents the methods to be used for the experimentation described in Section 4. Finally, Section 5 concludes the paper with some final remarks and points out future work.

2 Background

One of the parameters to be set when designing a human evaluation process is the number of items (i.e., either the number of questions an evaluator must answer or the number of tasks an evaluator must do) to obtain sufficiently reliable and representative results, while avoiding work overload. However, selecting a representative number of items is not a trivial task and depends on the type of study you are conducting.

In the field of statistics, there was a tendency to use the "n=30 rule-of-thumb" and set at least 30 as the default minimal number of questions or tasks for any study, but, to the best of our knowledge, without any scientific justification or empirical evidence. A possible explanation for this may have its origin in the pre-computer era, when all the calculations were made by hand. Student (1908) described how, when calculating the probable error of correlation coefficients, the best results were obtained with a sample size of 30 and one of the conclusions was "with samples of 30 [...] shows that the mean value approaches the real value [of the population] comparatively rapidly". Afterwards, the choice of this "magical" number as a sufficient sample size to get sounded results (from a statistical viewpoint) was maintained for decades, arguing that 30 samples were enough to hold the central limit theorem. Years later, in the computer era, this belief was deprecated in favor of bootstrap-based diagnostics (Hesterberg, 2008).

In the context of NLG human evaluation, Van der Lee et al. (2021) stated that "there should also be a sufficient number of outputs, so that a couple of particularly good or bad items do not skew the results too much. However, the number of items to evaluate depends heavily on the diversity of the sample, so we cannot give any specific recommendations here.". In their analysis of 89 papers, they noted that there was a median of 100 items utilized for human evaluation. However, the quantity of items varied widely, ranging from 2 to 5400, indicating a significant disparity. Thus, there is no general rule to determine the number of items that must be evaluated to obtain a reliable evaluation of an NLG system. Of course, the smaller the number of items required to get significant results the better. But beware of the negative oversimplification risk.

3 *Methods*

We aim to measure empirically the influence of the number of items when carrying out human evaluation of an NLG system. Considering that the number of items in our context refers to the number of texts to be evaluated when assessing the goodness of an NLG system, we will start from a large pool of texts evaluated by humans (what is taken as the baseline). Then, we will apply different resampling strategies in the search for the minimal set of texts taken from the initial pool that is required to draw sounded conclusions, i.e., to extract insights with a reasonable statistical significance. Before going in depth with the experimental study, let us introduce the resampling methods (see Section 3.1) and statistical tests (see Section 3.2) to be used later in Section 4.

3.1 Resampling methods

In statistics, resampling refers to the creation of new synthetic samples from observed or real ones. There are different methods to perform resampling such as cross-validation, permutation, subsampling or bootstrap. The latter is the one we used in our experiments and is introduced here to better understand the upcoming sections.

Bootstrap is an approach to statistical inference proposed by Efron (1979) which translates, in practice, the construction of different resampling schemes to approximate the sample distribution of a statistic (i.e., a function of the sample). The basic idea of bootstrap is that it is possible to make inferences about a given population from a reduced but representative sample of such population. If the target population were known, then we may measure the degree of agreement between the data distributions associated with the selected sample and with the entire population. Imagine that we desired to use answers to a survey to predict the result of a coming election in a city. If we may conduct the survey with all people who is entitled to vote, then we may have a high confidence in the predictive power of results of such survey. However, conducting a survey with the entire population is very expensive and sometimes even impossible because some people may refuse to take part. In practice, we should look for the smallest sample of the population that is representative enough of the entire population, so we can optimize resources and maximize the chance of getting significant results. However, what is the size of the smallest (but yet representative, for a certain significance) sample? Can we say that collecting answers to the survey by 30 people is enough? There is not a magical number a priori for the optimal sample size because the predictive power of the survey is not only a matter of quantity but also a matter of "inference" quality.

With the bootstrap method, the inferences are performed regarding synthetic samples, and they are derived by resampling from the given data which represents a subset of the target population. The true error in a sample statistic against its population is unknown because the entire population is unknown. However, the quality of inference of each synthetic sample from resampled data is measurable if we take as baseline the full initial sample (assuming it represents well the data distribution in the entire target population).

In short, the procedure to apply bootstrap is as follows:

1. Obtain a data sample that will be the "population" over which subsampling is applied.
2. Choose the number of synthetic subsamples ($replications$) to be generated.
3. Choose a subsample size (s) per $replication$.
4. For each $replication$:
   (a) Produce a new synthetic subsample with replacement of size s.
   (b) Estimate the quality of the generated subsample by computing the desired statistic.
5. Aggregate statistics for all $replications$.

It is worth noting that using the bootstrap method, for each replication, we always get a sample of the chosen size with replacement. For example, if the population includes 4 values such as [a,b,c,d] and we set 3 as subsample size, then we may obtain something like [a,a,c], where some values in the generated sample can be repeated. Thus, all values in the original population have the same probability to be selected when filling in each position in each new synthetic sample. On the contrary, if we applied Resampling without Replacement (RWOR), for each replication, then repetition of values is not allowed. In our experiment, we will test both bootstrap and RWOR.

3.2 ANOVA for discrete distributions

The analysis of variance (ANOVA) is a statistical test to compare the means of two different groups or populations (Fisher, 1992). ANOVA produces as output a number (named as F-statistic) and a p-value which supports or rejects the null hypothesis. In an ANOVA test, the null hypothesis is that the means of the groups being compared are the same, while the alternative hypothesis is that group means are different. This way, if the p-value obtained from the test is less than the usual α significance levels (0.1, 0.05, 0.01), then the null hypothesis can be rejected, and we can state that at least one of the means is different from the others. Then, different post-hoc tests can be applied to find out for which specific group the mean is different Otherwise, the null hypothesis cannot be rejected and therefore, we do not have enough evidence to say that there is a significant difference between the groups under comparison.

Even if ANOVA was originally defined for continuous data, there are some ANOVA extensions to treat properly also discrete data (De Leon and Zhu, 2008). In our case, we will use a variation of the ANOVA test which is more suitable to deal with discrete distributions. Namely, the function used is called `discANOVA`, from the WRS2 package in R (Mair and Wilcox, 2020). This function checks if the null hypothesis (i.e., that for two or more independent groups, the corresponding discrete distributions are identical) is satisfied. More precisely, `discANOVA` verifies if the groups have identical multinomial distributions.

It is worth noting that the power analysis done with software tools like G*Power (Faul et al., 2009) helps designers estimate the number of participants that are required in a user study with the aim of achieving significant results. However, as far as we know there is not any power analysis associated with the estimation of the number of items to assess by each participant.

4 *Experimentation*

For testing empirically the influence of the number of items in an NLG evaluation procedure, we followed the next steps: (i) we used a real NLG system to generate some texts (see Section 4.1); (ii) we proceeded with the human evaluation of all the generated texts (see Section 4.2); (iii) we created different prototypical evaluator profiles (see Section 4.3); and (iv) we tested the influence of the number of items for each of the evaluator profiles previously defined (see Sections 4.4 and 4.5).

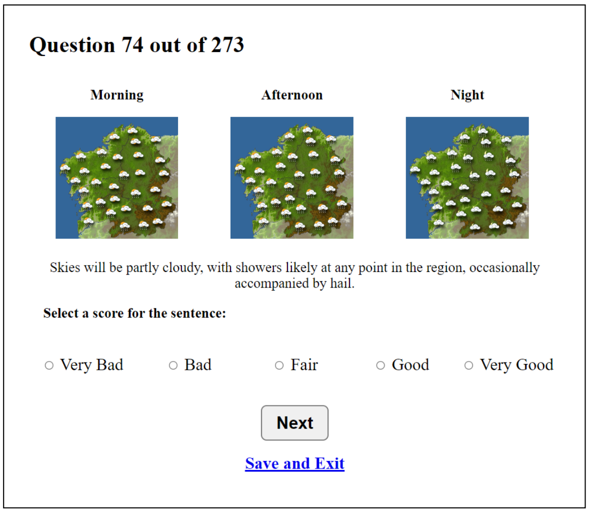


Figure 1: Example of question in the survey.

4.1 NLG system

We searched in the literature for a neural NLG system that may be used for generating the pool of texts to be evaluated, and we found that González Corbelle et al. (2022) released an NLG system along with the related dataset. Moreover, all the resources explained in the paper were available to reproduce the generation of texts.

The NLG system consists of a neural model which is adapted for a data to text (D2T) task in the context of meteorology. It generates short textual descriptions from meteorological tabular data. More precisely, it is an adaptation of a Transformer-based D2T model that initially was designed to generate chart captions (Obeid and Hoque, 2020). Regarding the dataset, we used the same as in the original model, composed of more than 3000 pairs of meteorological data and texts. We trained our model from scratch following the instructions given in the original paper (regarding the same parameters as well as the training, validation, and test partitions). As a result, we produced a total of 273 texts associated with the given test partition. These texts constitute the population to be evaluated in the following sections.

4.2 Human Evaluation

We designed an evaluation survey for assessing all the automatically generated texts. Each question from the survey was composed with a representation of the tabular input given to the system and the generated text.

The tabular data inputs represent the state of the sky for 32 different meteorological zones and 3 periods of the day (morning, afternoon, and night). Interpreting these 96-values table and

Score levels for evaluation	
Very bad	The description is not readable and does not match the data shown in the images, hallucinations are perceived in the generated text.
Bad	The description is not easy to read even though the content of the text is correct, but it ignores important information.
Fair	The description is readable, but not excessively natural. What is mentioned in the text is present in the data, but it is not complete enough.
Good	The description is well constructed, readable, and natural, but perhaps it could have mentioned some other relevant data present in the images.
Very Good	The description is so readable, natural, complete, and consistent with the data shown that could be considered a human text.

Table 1: Instructions given to the evaluators to score the texts based on their fluidity, naturalness, and content.

checking if the generated text describes the data correctly is a tedious task for an evaluator, so we opted for a simplified view of the input data. More precisely, we used the images available in the original data repository[1] instead of providing evaluators with raw data. For each question in the survey (see example in Figure 1), evaluators had to look at 3 meteorological maps (i.e., one for each period of the day) and rate how well (in a 5-point Likert scale from “Very Bad” to “Very Good”) the observed state of the sky is described by the given text (which was automatically generated by the D2T system). Before the evaluation, evaluators were given clear instructions about how to score the texts, based on their fluidity, correctness, and content (see Table 1).

[1] https://gitlab.citius.usc.es/gsi-nlg/meteogalicia-es

Annotators	**Cohen’s κ**	**Fleiss’ κ**
1 vs. 2	0.2128	
1 vs. 3	0.2565	0.2188
2 vs. 3	0.2119	

Table 2: Inter-Annotator Agreement coefficients: pair-wise Cohen’s Kappa and global Fleiss’ Kappa.

Three different evaluators with experience in the NLG field assessed the 273 texts generated by the system. Their Inter-Annotator Agreement was calculated using both the Cohen (1960) and Fleiss (1971) Kappa coefficients (see Table 2). Regarding the pair-wise agreement, i.e., the Cohen’s Kappa coefficient, we observe how the degree of agreement among the pairs of evaluators is similar in general (between 0.2 and 0.3). Nevertheless, going more into detail we could conclude that annotators 1 and 3 have the best agreement on their responses. Regarding the global agreement, i.e., the Fleiss’ Kappa coefficient, the reported value (0.2188) is in the same range of values reported by the Cohen’s coefficient. According to the Kappa statistic interpretation (Altman, 1991), both the pair-wise and global agreements are in the range $0.21 - 0.4$, that it is considered a “Fair Agreement”. However, the coefficients obtained are closer to the low part of the range, especially the Fleiss’ Kappa, which is far from the $0.41 - 0.6$ range considered as “Moderate Agreement” and even further from the “Good/Substantial Agreement” ($0.61 - 0.8$) that is deemed as desirable. This highlights the difficulty of the NLG evaluation task.

4.3 Evaluator profiles

For the sake of generality, we designed five different prototypical evaluator profiles taking as reference the responses collected in the previous survey. This was done in this way because we were looking for synthetic but realistic prototypical profiles.

On the one hand, evaluators with tendency to score high (“Good” or “Very Good”) most texts and not to penalize too much bad texts are considered to belong to a positive profile, while the evaluators whose tendency is just the opposite, i.e., to rate low (“Very Bad” or “Bad”) most texts, are considered as belonging to a negative profile. On the other hand, we can consider “bipolar" evaluators who tend to both extremes, i.e., only use very high and very low scores, or "neutral" evaluators who tend to "Fair" score for most cases. Bipolar evaluators are associated with a polarized profile, while neutral evaluators are associated with a neutral profile. In addition, there is a random profile which represents evaluators who vary randomly their scores in each question without any pre-defined criteria and do not fit in a specific evaluator profile of those already defined.

Considering these five evaluator profiles (positive, negative, neutral, polarized, and random), we proceed to generate their characteristic score distributions, by simulating as if an evaluator belonging to each of the profiles had evaluated all the cases under study. We take as a starting point the real scores collected in the previous survey and the generation procedure is made up of the following three steps:

1. **Transform all the responses from the three real human evaluators into three categories**: *negative*, *fair*, and *positive* responses.[2] Since we had five possible scores in a 5-point Likert scale, we aggregated all responses corresponding to "Very Bad" or "Bad" in the "negative" category, while responses associated to "Good" or "Very Good" go to the "positive" category. The "fair" category is made up of all responses with "Fair" scores.

2. **Aggregate the three evaluators' scores into a global curated score**: For each question, we apply the majority voting aggregation rule, i.e., if there are at least two evaluators that agree in the category of the response (negative, fair, positive), then such category defines the global score. Otherwise, the question is discarded from the aggregation process. As a result of the aggregation stage, we have a dataset with 246 curated cases, i.e., those cases in which there is at least a two-evaluators agreement.[3] The global distribution of scores is as follows: 89 cases are negative, 35 cases are fair, and 122 cases are positive (see Figure 2d).

3. **Create the five prototypical evaluator profiles**: We re-define 5-value distributions following the evaluation tendencies that are characteristic for each prototypical profile. From the dataset that we curated in the previous step, depending on the selected profile, we re-assign the cases into a different percentage for "Very Bad", "Bad", "Fair", "Good", and "Very Good" scores. Figure 2 shows the resultant score distribution for each evaluator profile, based on the profiles of the three human evaluators that were taken as reference. Figure 2d depicts the aggregated global scores from which the different evaluator profiles were generated. To do that, for each profile, the negative values are reassigned as "Very Bad" and "Bad" scores, while the positive values are reassigned as "Good" and "Very Good" scores. All the transformations are made in agreement with the expected tendencies for each evaluator profile. For example, if we look carefully at the distribution of cases in the Positive Profile (see Figure 2i) we can notice that from the 89 negative aggregated scores (see Figure 2d) 5% of cases are associated to the "Very Bad" score and 95% of cases are associated to the "Bad" score. Regarding the 122 positive values in the same picture, 50% of cases are associated to the "Good" score and the other 50% of cases is associated to the "Very Good" score. This way we produce a synthetic distribution of scores which is realistic (because it is grounded on the original human evaluations) but follows the expected tendency towards positive optimistic scores. Similar transformations produce the rest of profiles as depicted from Figure 2e to Figure 2h, always taking as starting point the aggregated global scores (Figure 2d) extracted from the real human evaluations.

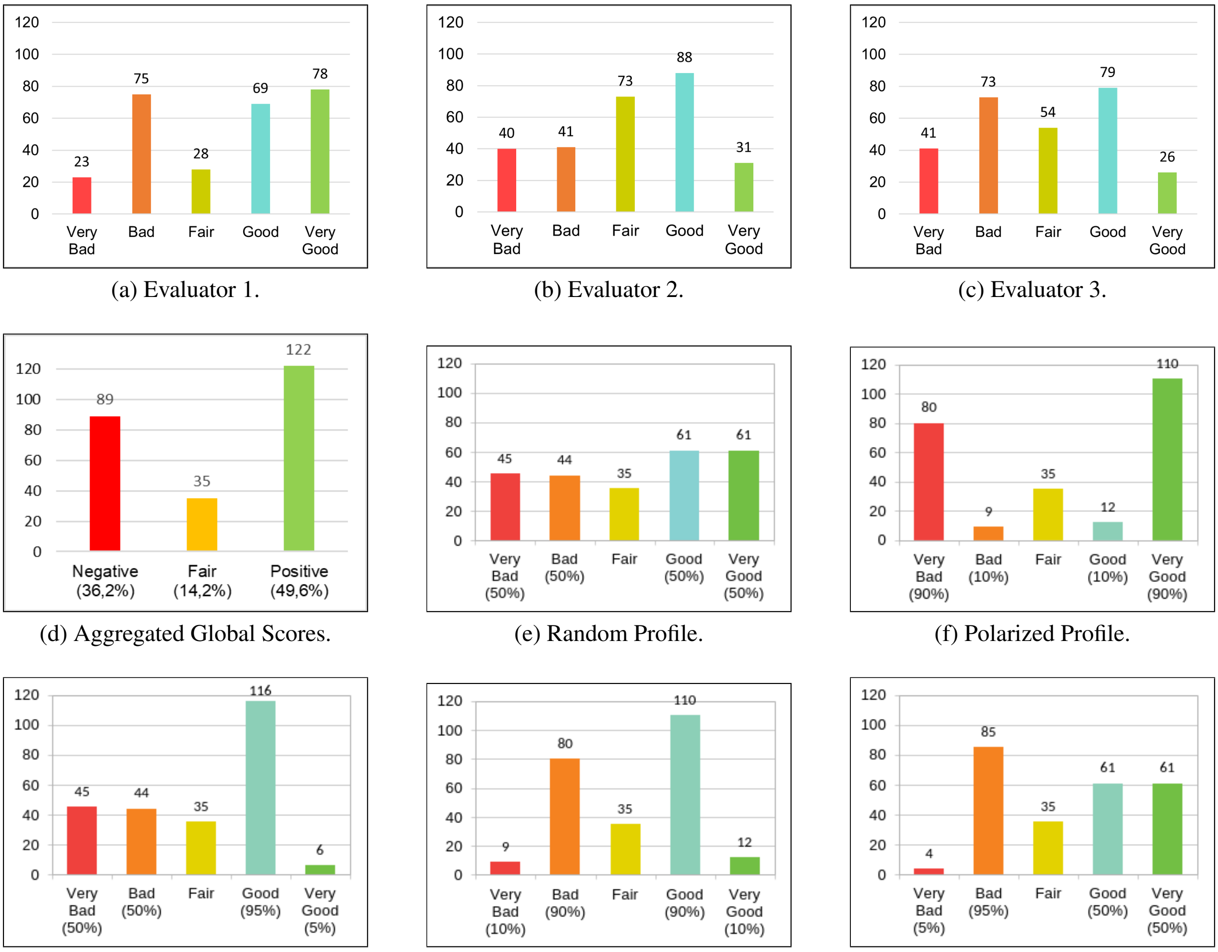
4.4 Resampling tests

In this section, we describe how to test the influence of the number of items in an NLG evaluation procedure. Considering the whole set of texts (S) generated by a system for a given test partition, if we evaluate a subset of texts $\tilde{S} \subset S$, some questions arise:

- How representative is $\tilde{S}$ (with respect to S)?
- Is the score distribution when evaluating $\tilde{S}$ the same as when evaluating S (considering the same evaluator or at least the same prototypical evaluator profile)?
- Which is the minimum number of samples in $\tilde{S}$ for yielding results as precise as the

[2] Note that here we are talking about response categories and not about evaluator profiles.

[3] Due to disagreement among evaluators, 27 cases were discarded.



(a) Evaluator 1. (b) Evaluator 2. (c) Evaluator 3.

(d) Aggregated Global Scores. (e) Random Profile. (f) Polarized Profile.

(g) Negative Profile. (h) Neutral Profile. (i) Positive Profile.

Figure 2: Original scores set by human evaluators (2a-2c), overall aggregated scores (2d), and generated score distributions for the synthetic prototypical evaluator profiles (2e-2i). The number below the bar labels is the percentage of negative/positive aggregated scores that were transformed into each level.

ones obtained with the whole set S?

- Does the number of items selected for $\tilde{S}$ have the same influence on different evaluators (or at least on different prototypical evaluator profiles)?

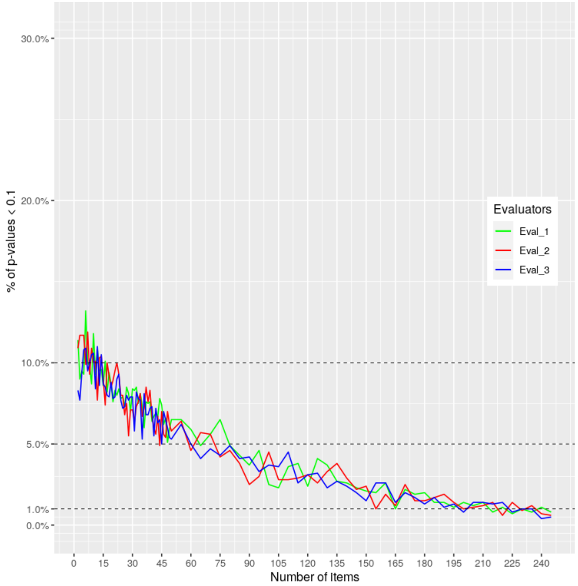
In the search for answers to the previous questions, the research hypothesis to validate is the following: we can approximate the “real” score distribution of an evaluator (i.e., the score obtained when such evaluator evaluates all texts in S) by evaluating only the items in $\tilde{S}$ and then applying a resampling method.

With the aim of testing if we can accept/reject the previous hypothesis, we apply bootstrap and RWOR resampling methods (as described in Section 3.1) on all the distributions shown in Figure 2. We set $\alpha = 0.1$. The number of replications is set to 1000. The sample size, which corresponds to the number of items to evaluate, ranges from 2 to 245. For each number of items tested in each of the distributions, we get 1000 p-values from the `discANOVA` output (i.e., one per replication). For each p-value, if it is lower than α, then we can say that in the given replication the resampled set of items $\tilde{S}$ has a distribution deemed as statistically different from the entire population S. It is worth noting that we count how many replications (out of 1000) yield to reject the null hypothesis: “the means of the groups S and $\tilde{S}$ are the same”.

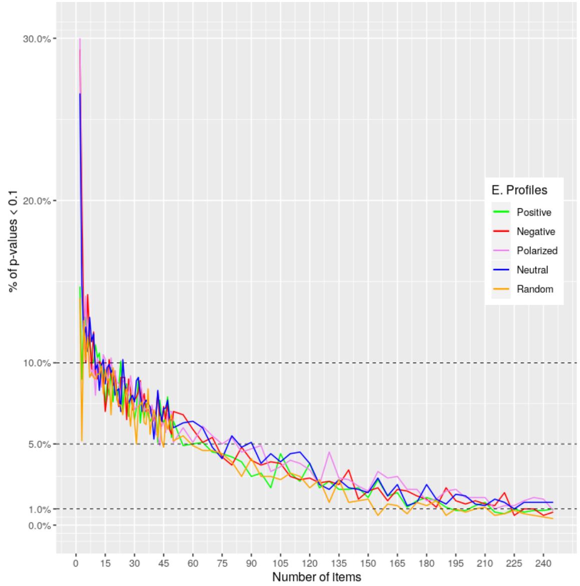
4.5 Results

Figure 3 summarizes the reported results. The comparison pays attention to the resampling methods (i.e., bootstrap and RWOR) and the type of evaluators (i.e., synthetic prototypical evaluator profiles vs. real evaluators).

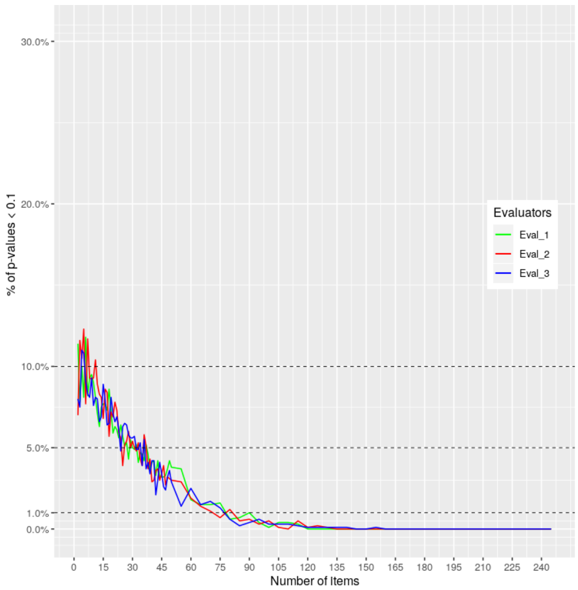
The general trend is that for a small number of items (i.e., less than 30) the hypothesis that



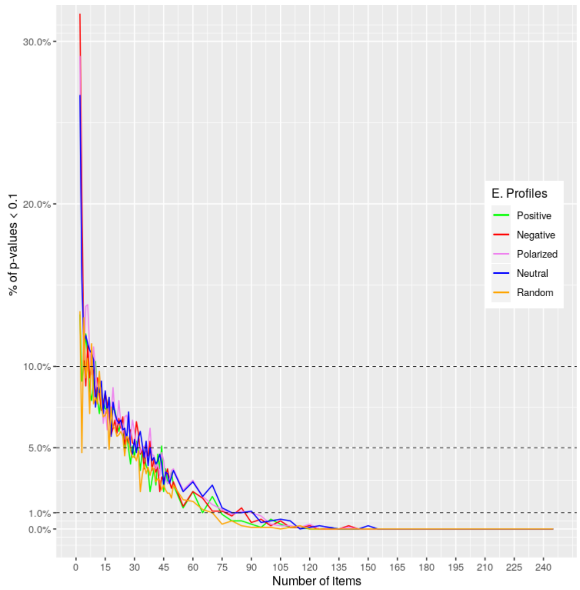
(a) Bootstrap: human evaluators.



(b) Bootstrap: prototypical evaluators.



(c) RWOR: human evaluators.



(d) RWOR: prototypical evaluators.

Figure 3: Results of running bootstrap and RWOR. For each number of items in the horizontal axis, 1000 samples were generated, and the picture shows the percentage of samples for which p-value < 0.1 when compared to the real distribution using `discANOVA`.

the two distributions are considered equal is rejected more often, no matter either the resampling method or the type of user. In the case of RWOR, values of 0% of hypothesis rejection are reached from 120 items on in some of the prototypical profiles (see Figure 3d), while for bootstrap the values tend asymptotically to 0% but never reach this value (see Figure 3b).

If we stick to reality, getting an exactly equal distribution from a small number of items is practically impossible. Therefore, we have established as a threshold of acceptance the cases in which less than 10% of the replications reject the hypothesis that the distributions are considered equal. Taking this threshold as a reference, we can see in Table 3 from what number of items less than 10% of cases reject the hypothesis, for each method and evaluator profile. The table also includes the minimal number of required items in case of establishing a smaller threshold value such as 5% or 1%. It is easy to appreciate how the smaller the pre-defined threshold, the bigger the number of items that are required to get significant results.

	Bootstrap			**RWOR**		
Threshold	***10%***	***5%***	***1%***	***10%***	***5%***	***1%***
Eval_1	16	75	240	6	37	75
Eval_2	22	70	235	11	37	80
Eval_3	14	55	220	5	36	75
Positive	23	65	210	10	44	70
Negative	17	70	220	8	38	85
Polarized	18	80	240	9	38	85
Neutral	24	90	245	9	36	90
Random	18	55	210	8	27	65

Table 3: For each human evaluator and for each prototypical evaluator profile, number of items from which the % of samples with a p-value $<$ 0.1 is always lower than a pre-defined threshold as illustrated in Figure 3 (10%, 5%, 1%).

If we look at the bootstrap method, the evaluator that first reaches the threshold of 10% of rejection is the third one (Eval_3 in Figure 3a) with 14 items. In addition, the prototypical profiles that need the most items to achieve a distribution equivalent to the original one are the Neutral and Positive profiles, with 24 and 23 items, respectively (see Figure 3b). This is not the case for RWOR (see Figure 3c) which yields much lower numbers, with the distribution of Eval_3 corresponding to the lowest value (5 items), while Eval_2 is associated with the highest value (11 items).

On the one hand, taking the bootstrap method as a reference, with a rejection threshold of 10%, we could say that for any of the tested evaluators and prototypical profiles, from at least 24 items, we could obtain a distribution of scores equivalent to performing the complete evaluation of all the curated 246 test cases. On the other hand, if we consider the RWOR method for the same threshold, the number of items to obtain a distribution equivalent to the original one is reduced to no more than 11, for all evaluators and prototypical profiles tested in this experiment.

To sum up with, reported results validate our research hypothesis for the specific experimental setting under consideration, i.e., thanks to the use of resampling methods we can get sounded evaluation insights while requiring a very small number of items to be evaluated. The actual numbers are detailed in Table 3. For a 10% threshold, in the worst case, only 9.75% of the items in S needs to be evaluated. In the best case, the percentage of items to evaluate is only 2%. If a smaller threshold were required, then the number of items should be bigger. Moreover, the number of items required by RWOR is always much smaller. In the worst case, (i.e., the Neutral prototypical profile with threshold value equal to 1%) the required number of items is 36.58% of all the items in S.

5 Final Remarks and Future Work

In this paper we tested the influence of the number of items in a human evaluation of NLG systems. To do so, we first carried out an evaluation with three different raters on a pool of texts generated by a Data-To-Text neural system. Then, with the scores obtained from the evaluation of all the texts, we created different prototypical evaluator profiles (that are synthetic but realistic because they are grounded on the previous human evaluations). Finally, using resampling methods, we simulated evaluations in the search for the minimal number of items that is required to get sounded insights.

After carrying out the experimentation and analyzing the results obtained, we can conclude that in our case is possible to approximate the distribution of evaluations of a real set of texts from a smaller subset of evaluated items. In our experiment, with a test set of 246 items and each text evaluated in a 5-point Likert scale, it would be sufficient to evaluate 24 items (i.e., about 10% of items randomly taken from the entire pool of texts) to ensure that, no matter the prototypical evaluator profile, we obtain a score distribution equivalent to evaluating all the texts generated by the system in at least 90% of the cases. This fact validates the research hypothesis under study: "well-known resampling statistical methods can contribute to get significant results even with a small number of items to be evaluated by each evaluator".

Regarding the already mentioned "n=30 rule-of-thumb" we can say that for the specific experimental setting we achieved good results even without reaching 30 items in the evaluation. Nonetheless, the interesting finding is that for different evaluators and prototypical profiles this number varies and seems that it is not possible to have an ideal number of items for all evaluations beforehand. Considering evaluators with different profiles may mean that approximating the actual distribution of scores requires a higher/lower number of items to be evaluated. Moreover, the minimal number of items depends also on the pre-defined threshold. Thus, the ideal number of items to obtain reliable results in an NLG evaluation cannot be generalized.

Anyway, our empirical study represents a step forward in the search for an evaluation protocol admitted worldwide. The empirical results highlight the importance of carefully addressing the experimental setting in human evaluation studies for NLG systems. It is crucial to pay special attention to those parameters chosen in the context of the evaluation process, being the number of items especially relevant because it can reduce dramatically the evaluation costs if it is properly selected. In addition, this work provides readers with a benchmark for choosing the ideal number of items for a given evaluation study, since all related resources are available online as open access.[4]

As future work, we plan to extend the empirical study to other types of evaluations in which the scoring criteria and scale may vary from those tested in this work. Also, alternative approaches or formulas for calculating and determining the minimum required number of items to achieve representative results from a sample will be examined. Moreover, we will consider how resampling methods can be integrated in the evaluation procedure to address the lack of resources (e.g., evaluators availability) in NLG human evaluation.

Acknowledgments

J. González-Corbelle, J.M. Alonso-Moral and A. Bugarín-Diz acknowledge the support from the Galician Ministry of Culture, Education, Professional Training and University (grants ED431G2019/04 and ED431C2022/19). These grants are co-funded by the European Regional Development Fund (ERDF/FEDER program). In addition, this work is supported by Grants PID2021-123152OB-C21 and PID2020-112623GB-I00 funded by MCIN/AEI/10.13039/501100011033 and by "ESF Investing in your future", and by Grant TED2021-130295B-C33 funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR". R.M. Crujeiras acknowledges the support from project PID2020-116587GB-I00, funded by MCIN/AEI/10.13039/501100011033 and the Competitive Reference Groups 2021-2024 (ED431C 2021/24) from the Xunta de Galicia.

[4]`https://gitlab.citius.usc.es/gsi-nlg/human-evaluation-resampling`

References

Altman, D. G. 1991. *Practical Statistics for Medical Research*. Chapman and Hall.

Banerjee, S. and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Belz, A. 2022. A Metrological Perspective on Reproducibility in NLP. *Computational Linguistics*, 48(4):1125–1135.

Belz, A., C. Thomson, E. Reiter, and S. Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

De Leon, A. and Y. Zhu. 2008. ANOVA extensions for mixed discrete and continuous data. *Computational Statistics Data Analysis*, 52(4):2218–2227.

Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

Faul, F., E. Erdfelder, A. Buchner, and A.-G. Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41:1149–1160.

Fisher, R. A., 1992. *Breakthroughs in Statistics: Methodology and Distribution*, chapter Statistical Methods for Research Workers, pages 66–70. Springer New York, New York, NY.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

González Corbelle, J., A. Bugarín-Diz, J. Alonso-Moral, and J. Taboada. 2022. Dealing with hallucination and omission in neural natural language generation: A use case on meteorology. In *Proceedings of the 15th International Conference on Natural*

Language Generation, pages 121–130, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Hesterberg, T. 2008. It's time to retire the "n $\geqslant$ 30" rule. In *Proceedings of the American Statistical Association*, Alexandria VA.

Kane, H., M. Y. Kocyigit, A. Abdalla, P. Ajanoh, and M. Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In S. Agarwal, O. Dušek, S. Gehrmann, D. Gkatzia, I. Konstas, E. Van Miltenburg, and S. Santhanam, editors, *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland). Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Mair, P. and R. Wilcox. 2020. Robust Statistical Methods in R Using the WRS2 Package. *Behavior Research Methods*, 52:464–488.

Moramarco, F., A. Papadopoulos Korfiatis, M. Perera, D. Juric, J. Flann, E. Reiter, A. Belz, and A. Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Obeid, J. and E. Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the Transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland. Association for Computational Linguistics.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 311–318, USA. Association for Computational Linguistics.

Reiter, E. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Sellam, T., D. Das, and A. Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. Association for Computational Linguistics.

Student. 1908. Probable error of a correlation coefficient. *Biometrika*, 6(2/3):302–310.

Van der Lee, C., A. Gatt, E. van Miltenburg, and E. Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151.1–101151.24.

Wang, J., Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, and J. Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In Y. Dong, W. Xiao, L. Wang, F. Liu, and G. Carenini, editors, *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Hybrid. Association for Computational Linguistics.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*. OpenReview.

Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 57-71 recibido 13-12-2023 revisado 13-01-2024 aceptado 28-01-2024

Introducing the NLP task of negative attitudinal function identification

Introduciendo la tarea de PLN de identificación de funciones actitudinales negativas

Nicolás José Fernández-Martínez
Departamento de Filología Inglesa, Universidad de Jaén
njfernan@ujaen.es

Abstract: On social media, users often express emotions, judgments, and evaluations on various social and private topics, detectable through automated methods. While NLP tasks like emotion detection and dialogue act classification focus on identifying emotions and intentions in texts, little attention has been paid to the attitudinal function of a text, such as expressing dislike, disagreement, pessimism, disapproval, etc. Our main contribution introduces the NLP task of negative attitudinal function identification, going beyond emotion detection and dialogue classification by focusing on users' intent and the expression of negative emotions, and negative ethical and aesthetic evaluations. We present a basic synthetic dataset for negative attitudinal functions built with foreign language teaching and learning resources. The dataset was used to develop negative attitudinal function models with supervised approaches, which were compared against other baseline models based on social media emotion detection datasets whose emotion categories were mapped to negative attitudinal functions. Our models, though not consistently outperforming baselines due to the qualitative differences of the tasks, use of out-of-domain data, and labeling issues found in the emotion detection datasets, exhibit promising capabilities with unseen data and in multilingual contexts.
Keywords: negative attitudinal function identification, NLP task, social media, synthetic dataset.

Resumen: En las redes sociales, los usuarios expresan con frecuencia sus emociones, juicios y evaluaciones sobre diversos temas sociales y privados, detectables mediante métodos automatizados. Mientras que tareas de PLN como la detección de emociones y la clasificación de actos de diálogo se centran en identificar emociones e intenciones en los textos, se ha prestado poca atención a la función actitudinal de un texto, como expresar desagrado, desacuerdo, pesimismo, desaprobación, etc. Nuestra principal contribución introduce la tarea de PLN de identificación de funciones actitudinales negativas, yendo más allá de la detección de emociones y la clasificación de diálogos al centrarse en la intención de los usuarios y la expresión de emociones negativas y evaluaciones éticas y estéticas negativas. Presentamos un dataset sintético básico para funciones actitudinales negativas construido con recursos obtenidos del campo de la enseñanza y aprendizaje de lenguas extranjeras. El conjunto de datos se utilizó para desarrollar modelos supervisados de funciones actitudinales negativas, que se comparó con otros modelos estándar basados en datasets de detección de emociones de redes sociales cuyas categorías de emociones fueron reasignadas a funciones actitudinales negativas. Nuestros modelos, aunque no superan sistemáticamente los modelos estándar debido a las diferencias cualitativas de las tareas, el uso de datos fuera de dominio y los problemas de etiquetado encontrados en los datasets de detección de emociones, muestran capacidades prometedoras con datos nunca antes vistos y en contextos multilingües.
Palabras clave: identificación de funciones actitudinales negativas, tarea de PLN, redes sociales, dataset sintético.

ISSN 1135-5948 DOI 10.26342/2024-72-4

1 Introduction

On social media platforms such as Twitter (now called X), which has an ever-growing active user base of around 450 million monthly users, users post a staggering 500 million tweets a day on a wide range of public and private issues(Ruby, 2023).[1] This wealth of digital data can be leveraged with Natural Language Processing (NLP) tasks that focus on detecting users' subjective emotions (i.e. emotion detection) (Cambria, 2016) or intent (i.e. dialog act classification) (Jurafsky et al., 1998). These NLP tasks provide manifold practical applications across diverse domains, from business and commerce to politics, disaster management, sociology, and digital humanities (Mohammad, 2021; Jurafsky and Martin, 2023).

Despite advances in understanding emotions and intent, a critical gap in existing NLP tasks remains unaddressed — the identification of the attitudinal function(s) of a given message. In linguistic terms, an attitudinal function reflects the speaker's intention to communicate emotions, judgments, and appreciations about a topic, entity, or thing. This is particularly relevant in the context of smart cities, where predicting citizens' dissatisfaction and concerns plays a pivotal role in enhancing their well-being and improving city services and infrastructure (Periñán-Pascual, 2023). The focus on negative attitudinal functions in smart city contexts results from the research conducted in the ALLEGRO project (Periñán-Pascual, 2023) where smart city problems are identified. Negative attitudinal functions take center stage in smart city contexts, reflecting citizens' dissatisfaction with various aspects of their lives, from the condition of streets and parks to broader sociological issues such as economic inequality, racism, sexism, and political concerns. These functions encompass a spectrum of semantic subtleties, including expressions of dislike, disagreement, indifference, anger, threats, worries, distrust, pessimism, and more.

To bridge this gap, we reuse the concept of attitudinal function from linguistic theory. Drawing inspiration from functional theories of foreign language teaching and learning, systemic functional linguistics, and speech act theory, we introduce the novel NLP task of negative attitudinal function identification. This task extends beyond emotion detection and dialog act classification, with a linguistically informed set of categories for dealing with the expression of intent with respect to emotions, judgements, and evaluations. We developed a basic synthetic dataset that captures negative attitudinal functions in smart city scenarios using existing linguistic resources. This dataset contains prototypical lexico-grammatical patterns that are formally realized by functions.[2] Leveraging this dataset, various supervised models were developed, including fine-tuned Transformers, contextualized sentence embeddings, and zero-shot classification with Natural Language Inference (NLI) using Transformers, together with traditional Machine Learning using a Naïve Bayes model.

Our study explores the models' efficacy in identifying negative attitudinal functions in multilingual contexts, as well as the potential reuse of emotion detection datasets for this task. Our research seeks to contribute to the evolving landscape of NLP by addressing nuanced aspects of user expressions that were not captured in emotion detection and dialog act classification. This manuscript is organized as follows. Section 2 provides the background on emotion detection, dialog act classification, and negative attitudinal function identification. Section 3 describes the methodology used in our experiment, including the development of the synthetic dataset of negative attitudinal functions, the automatic mapping of social media emotion detection datasets, and the supervised approaches. Section 4 gives the results and discussion, including limitations, challenges, and future research directions. Section 5 presents the conclusion.

2 Background

2.1 Emotion detection

Emotions are integral to human life, influencing communication, interaction, and learning (Scherer, 2005). Studied historically in philosophy, particularly in Socratic schools like Aristotelianism and Stoicism (Sorabji, 2002),

[1] https://www.demandsage.com/twitter-statistics/

[2] The dataset is available under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license at the following link: https://github.com/njfm0001/Negative-Attitudinal-Function-Identification

emotions are now being addressed in the psychological and computational literature (Picard, 1997; Scherer, 2005). Emotion detection in NLP is a sequence classification task (Cambria, 2016; Mohammad, 2021). Categorical models often rely on Ekman (1992)'s six universal emotions or Plutchik (1980)'s extended model. Applications span public health, safety, crisis response, advertising, and entertainment (Cambria, 2016; Mohammad, 2021). Users express emotions in text through emotionally loaded words, emojis, or implicit cultural cues (Mohammad, 2021). Important emotion markers include emotion verbs, adjectives, orthography, terms of address, adverbs, and exclamations (Mohammad and Alm, 2015; Alba-Juez, 2018). Emotions in text can refer to the writer, reader, or characters (Maia and Santos, 2018), although they typically correspond to the writer's emotion (Buechel and Hahn, 2017). This task faces challenges like figurative language and cultural connotations (Mohammad, 2021). Efforts are now focusing on semantic role labeling in appraisal-based approaches (Campagnano, Conia, and Navigli, 2022; Wegge et al., 2022; Troiano, Oberländer, and Klinger, 2023).

Datasets for emotion detection come from various genres or domains, such as interviews (Scherer and Wallbott, 1994), news headlines (Strapparava and Mihalcea, 2007; Bostan, Kim, and Klinger, 2020), blog posts (Aman and Szpakowicz, 2007; Quan and Ren, 2009), tales (Alm, Roth, and Sproat, 2005), and, more recently, tweets (Mohammad, 2012; Roberts et al., 2012; Mohammad et al., 2015; Liew, Turtle, and Liddy, 2016; Mohammad et al., 2018; Plaza del Arco et al., 2020; Saravia et al., 2018) and Reddit comments (Demszky et al., 2020; Dwivedi-Yu and Halevy, 2022). Each domain has its own linguistic idiosyncrasies (Bostan and Klinger, 2018). Annotation methods include self-reported interviews (Scherer and Wallbott, 1994), distant supervision (Mohammad, 2012; Purver and Battersby, 2012; Wang et al., 2012; Dwivedi-Yu and Halevy, 2022), expert knowledge (Aman and Szpakowicz, 2007; Strapparava and Mihalcea, 2007), and crowd-sourcing (Mohammad and Alm, 2015; Liew, Turtle, and Liddy, 2016; Schuff et al., 2017; Demszky et al., 2020; Plaza del Arco et al., 2020). Approaches to emotion detection include symbolic models with lexica and rule-based methods (Strapparava and Mihalcea, 2007; Dini and Bittar, 2016; Semeraro et al., 2023), and probabilistic models using machine learning or deep learning (Aman and Szpakowicz, 2007; Mohammad, 2012; Liew and Turtle, 2016). Probabilistic, deep learning models use networks like bidirectional LSTM or Transformers such as BERT (Devlin et al., 2019) or XLM-Twitter (Barbieri, Anke, and Camacho-Collados, 2022) via fine-tuning (Demszky et al., 2020; Vera, Araque, and Iglesias, 2021; Dwivedi-Yu and Halevy, 2022; Aroyehun et al., 2023) or through zero-shot classifiers (Basile, Pérez-Torró, and Franco-Salvador, 2021; Plaza del Arco, Martín-Valdivia, and Klinger, 2022; Yang et al., 2023).

2.2 Dialog act classification

Dialog act classification involves identifying users' communicative intent in texts, typically dialogs (Searle, 1976; Jurafsky and Martin, 2023). Various annotation schemes like DAMSL (Core and Allen, 1997), SWBDD (Jurafsky, Shriberg, and Biasca, 1997), SPAAC (Leech and Weisser, 2003), DiAML (Bunt, 2009), DART (Weisser, 2018), and MIDAS (Yu and Yu, 2021) differ in dimensions, domains, and segmentation levels. DAMSL, the first scheme, is multidimensional, while SWBDD simplifies it for tagging dialog units. DiAML introduces DIT++ for annotation, encompassing clusters of general-purpose and dimension-specific functions. DART offers a fine-grained classification of speech acts. MIDAS is a multi-label scheme for human-machine interaction. These schemes apply to multi-turn texts but not tweets or other genres (Bunt, 2019).

Tagging pragmatic phenomena related to user intent faces challenges because of the complexity of the tagging schemes, requiring expert knowledge and labor-intensive efforts (Weisser, 2018; Yu et al., 2023). Datasets like Switchboard Dialog Act Corpus (Jurafsky, Shriberg, and Biasca, 1997), DailyDialog (Li et al., 2017), and DialogBank (Bunt et al., 2016) use various tagging schemes. Social media datasets, especially Twitter (now X), present difficulties due to their idiosyncratic nature (Saha et al., 2020; Baldwin et al., 2013). Symbolic approaches use lexico-syntactic cues like performative verbs and punctuation marks (Jurafsky et al., 1998).

Probabilistic models include Hidden Markov Models, Support Vector Machines, or Logistic Regression (Stolcke et al., 2000; Zhang, Gao, and Li, 2011; Vosoughi and Roy, 2016), and deep learning models include LSTM networks and Transformers (Khanpour, Guntakandla, and Nielsen, 2016; Cerisara et al., 2018; Enayet and Sukthankar, 2020; Saha et al., 2020; Żelasko, Pappagari, and Dehak, 2021; Gung et al., 2023; Ostyakova et al., 2023a).

2.3 Negative attitudinal function identification

The appraisal framework in systemic functional linguistics explores three semantic domains related to the interpersonal metafunction: attitude, engagement, and graduation (Martin and White, 2005; Bednarek, 2008). Our focus is on attitude, which comprises affect (emotion) and opinion (judgment and appreciation). Affect pertains to emotional reactions, judgment involves ethical evaluations, and appreciation centers on aesthetic evaluations. In speech act theory (Searle, 1976), expressive and emotive speech acts are relevant to attitudes, dealing with psychological states (Sbisà, 1975; Guiraud et al., 2011; Zhabotynska and Slyvka, 2020). Commisive and directive speech acts also carry attitudinal overtones (Sbisà, 1975). Existing speech act tagging schemes such as DiAML and DART offer some insights into expressing attitudes through speech acts. However, finer-grained distinctions are lacking.

To broaden the understanding of attitudinal functions, we also draw on the concept of communicative function from foreign language teaching and learning studies (Finocchiaro and Brumfit, 1983; Milanovic and Saville, 2012). Communicative functions serve specific purposes, including expressing attitudes. We developed a taxonomy of negative attitudinal functions for smart city scenarios, considering linguistic insights from all the previous approaches and function categorizations (Blundell, Higgens, and Middlemiss, 1982; Wilkins, 1976). The reason to focus on negative attitudinal functions is because, in smart city contexts, attitudes are often expressed in relation to negative events or entities, reflecting citizens' feelings and ethical and aesthetic evaluations. For example, people might communicate their feelings towards inflation (affect), passing judgments on economic policymakers for their decisions (judgment), or evaluating the effectiveness of economic policies independently of the policymakers' actions (appreciation).

Our approach to attitudinal functions is a formal and discourse-pragmatic one, encompassing affect and evaluation through judgments and appreciations, and involves developing a dataset of attitudinal constructions that captures attitudinal meanings through lexico-grammatical patterns or constructions.

2.3.1 The task of negative attitudinal function identification

We define negative attitudinal function identification as a sequence classification task, where a given word sequence s is assigned one or more negative attitudinal function labels f from a set $f_1, f_2, f_3, \ldots$, making it a single-label or multi-label task.

While linguistic constructions in the taxonomy serve as the primary cue for function identification, implicit expressions may also exist through other means (e.g. emojis, hashtags). For instance, the lexico-syntactic cue *disagree with something* clearly indicates the attitudinal function DISAGREE. However, in complex microtexts, multiple functions may be implied, requiring attitudinal function models to discern nuances. Defining functions by form poses challenges, as expressions like *I really hate _* may refer to multiple functions simultaneously (i.e. DISLIKE and ANGER). In social media microtexts, determining attitudinal functions becomes intricate due to potential implicit markers and polysemy.

Our synthetic dataset serves as a starting point for annotators or contextually rich models, like Transformers, to recognize the subtleties in expressing attitudinal functions, considering the complexities of context and implicit markers.

2.3.2 Differences and similarities with respect to emotion detection and dialog act classification

All these tasks involve sequence or text classification, assigning predefined categories to given texts, but they differ in aims, scope, and categories.

In emotion detection, the goal is to infer the writer's, reader's, and/or characters' emotional states (Picard, 1997) using, for instance, categorical models based on motivational theories (Ekman, 1992; Plutchik, 1980). The focus is thus on a set of universal emotions, employing psychological categorizations. Ekman (1992)'s categorization comprises the following emotion categories: anger, disgust, fear, joy, sadness, surprise. Plutchik (1980)'s tagset adds anticipation and trust.

Dialog act classification aims to identify users' intent in utterances (usually in dialogs) using dialog acts (Jurafsky and Martin, 2023), loosely based on speech acts (Searle, 1976). Widely used dialog act tagsets include statements, questions, suggestions, comments, and miscellanea (Zhang, Gao, and Li, 2011). While most dialog acts lack attitudinal meanings, some taxonomies like DIT++ (Bunt, 2009) and DART (Weisser, 2018) include some attitudinal aspects that have not been explored in practice.

Negative attitudinal function identification detects users' negative attitudinal intent, addressing emotional states, reactions, evaluative meanings (judgments and appreciations), and subjective states beyond emotional meanings. It bridges the intent and emotion focus of dialog act classification and emotion detection, encompassing linguistic nuances like distinctions between emotion, judgment, and appreciation. Unlike emotion detection, it adopts a linguistically grounded approach to cover a broader range of attitudinal meanings on the basis of categorizations found in theoretical and applied linguistic theory.

3 Methods

3.1 Development of the synthetic dataset of negative attitudinal functions

In our experimental setup, we built a synthetic dataset of linguistic constructions labeled with negative attitudinal functions, drawing from Blundell, Higgens, and Middlemiss (1982)'s comprehensive list. Blundell, Higgens, and Middlemiss (1982) is a practical textbook of functions for teaching and learning English as a Foreign Language that contains communicative functions of the following types: informational, attitudinal, and active. There are 12 informational functions, related to seeking or providing information; 48 attitudinal functions, in which an attitude is expressed towards something (e.g. feelings, opinions, judgments); and 32 active functions, to establish courses of actions. Other function categories are given related to social formulas, communication strategies, and metalinguistic questions. This list was largely based on earlier categorizations in the functional-notional approach to foreign language teaching (Wilkins, 1976; Finocchiaro and Brumfit, 1983). Focusing on those attitudinal functions with negative overtones, we manually selected categories that were most relevant to smart cities and social media, where users express negative attitudes. Categories include PESSIMISTIC, WORRIED, ANGRY, DISAPPOINTED, BORED, DISLIKE, NOT_APPROVE, NOT_IMPORTANT, NOT_INTERESTED, DISAGREE, NOT_CORRECT, WARN, COMPLAIN, THREATEN, UNWILLING, REFUSE. Additionally, we introduced DISTRUST, using the Collins dictionary and the MacMillan dictionary thesauri, because this category was deemed crucial for smart city contexts, and the OTHER category, containing constructions from functions of different types found in Blundell, Higgens, and Middlemiss (1982). This results in 18 attitudinal functions with 362 constructions (Table 1).

Label	**No.**
OTHER	79
NOT_CORRECT	8
PESSIMISTIC	18
WORRIED	16
ANGRY	23
DISAPPOINTED	8
BORED	18
DISLIKE	22
NOT_APPROVE	19
NOT_IMPORTANT	20
NOT_INTERESTED	17
DISAGREE	23
WARN	10
COMPLAIN	21
THREATEN	13
UNWILLING	13
REFUSE	19
DISTRUST	15
Total	**362**

Table 1: Taxonomy of negative attitudinal functions.

Label	Examples
OTHER	My pleasure.
	Don't worry (about _).
NOT_CORRECT	You're/He's/She's/That's/We're/They're (all) wrong.
	(That's/It's) nonsense/rubbish/bullshit/bs/crap.
PESSIMISTIC	(I'm) not (too) happy (about _).
	There's no way.
WORRIED	I fear _.
	(I'm) (very) worried/uneasy (about _).
ANGRY	What an idiot/fool.
	_ (really) makes me mad.
DISAPPOINTED	That's/It's a real shame/pity/let-down.
	What a pity/disappointment.
BORED	_ is a (total) bore/drag.
	_ leaves me cold.
DISLIKE	(I) don't like _.
	How awful.
NOT_APPROVE	(I) don't think that's/it's (very) good.
	(I'm) dead against _.
NOT_IMPORTANT	(I) don't think that's (so) important.
	Does _ matter?
NOT_INTERESTED	(I'm) not (very) interested (in _).
	(I) couldn't care less (about _).
DISAGREE	(I) don't agree (with _).
	(I) can't go along (with _).
WARN	Watch out (for _).
	Make sure you don't do _.
COMPLAIN	_ really is the limit!
	(I'm) not at all satisfied (with _).
THREATEN	If I were you, I wouldn't do _.
	Don't do that or I'll do _.
UNWILLING	(I) don't (really) fancy doing _.
	I'd rather not (do _).
REFUSE	(I'm) sorry, I can't/couldn't (do _).
	Out of the question.
DISTRUST	Are you kidding?
	You must be joking.

Table 2: Examples of negative attitudinal function constructions.

The dataset construction is linguistically and sociologically motivated: it is informed by expert linguistic knowledge and sociological insights into citizens' problems in smart city contexts (Periñán-Pascual, 2023). Many constructions were taken literally from Blundell, Higgens, and Middlemiss (1982), whereas others were adapted to fit the linguistic nature of the social media domain. The constructions follow methodological conventions involving parentheses, underscores, and slashes, for optional elements, unspecified topics, and alternative expressions, respectively. Rules apply to the 362 constructions to obtain the full synthetic dataset with 902 samples. For example, for the COMPLAINT category, the constructions *(I'm) not at all satisfied (with _)*, after the application of the rules, generate the samples *not at all satisfied*, *not at all satisfied with it*, *I'm not at all satisfied*, and *I'm not at all satisfied with it.* Another example: for the NOT_CORRECT category, the construction *(That's/It's) nonsense/rubbish/bullshit/bs/crap.* generates the samples *nonsense*, *rubbish*, *bullshit*, *bs*, *crap*, *That's nonsense*, *That's rubbish*, *That's bullshit*, *That's bs*, *That's crap*, *It's nonsense*, *It's rubbish*, *It's bullshit*, *It's bs*, and *It's crap.* Some other examples can be found in Table 2.

3.2 Automatic mapping of social media emotion detection datasets

Our experiment utilized state-of-the-art datasets from emotion detection tasks, as some attitudinal functions involve explicit or implicit emotion expressions. Specifically, we selected social media datasets due to their relevance in expressing attitudes of negative type. These include EmoEvent (Plaza del Arco et al., 2020), GoEmotions (Demszky et al., 2020), CARER (Saravia et al., 2018), and the AIT dataset (Mohammad et al., 2018). For multilingual datasets (e.g. EmoEvent with English and Spanish tweets or AIT with English, Spanish, and Arabic tweets), we merged them while respecting the original splits and without pre-processing. Each dataset is briefly described below:

- **EmoEvent**: it comprises 8,409 English and 7,303 Spanish tweets labeled with a single emotion category, topic, and offensiveness (Plaza del Arco et al., 2020). The emotion set aligns with Ekman (1992) (i.e. anger, disgust, fear, joy, sadness, surprise), including the 'other' category.
- **GoEmotion**s: it consists of around 54,000 English Reddit comments annotated with 27 fine-grained emotions plus a neutral category (Demszky et al., 2020). The dataset is highly imbalanced.
- **CARER**: A dataset of tweets expressing emotions, loosely based on Ekman (1992)'s classification (i.e. anger, fear, joy, love, sadness, surprise) (Saravia et al., 2018). Built via distant supervision using emotion-related hashtags.
- **AIT dataset**: Used for SemEval-2018 Task 1, it includes English, Spanish, and Arabic tweets (Mohammad and Kiritchenko, 2018). Emotion categories are based on Plutchik (1980)'s classification: anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust.

The mapping of emotions to negative attitudinal functions considered the different annotation schemes. The mapping process is detailed in Table 3. The negative attitudinal function identification task was evaluated as either single-label or multi-label depending on the tagging scheme of the dataset.

Dataset	Mapping
EmoEvent	ANGRY: anger DISLIKE: disgust WORRIED: fear PESSIMISTIC: sadness OTHER: joy, surprise, other
GoEmotions	OTHER: admiration, amusement, approval, caring, confusion, curiosity, desire, embarrassment, excitement, gratitude, joy, love, optimism, pride, realization, relief, surprise, neutral ANGRY: anger DISLIKE: annoyance, disgust DISAPPOINTED: disappointment NOT_APPROVE: disapproval WORRIED: fear PESSIMISTIC: grief, remorse, sadness
CARER	ANGRY: anger WORRIED: fear PESSIMISTIC: sadness OTHER: joy, love, surprise
AIT	ANGRY: anger DISLIKE: disgust WORRIED: fear PESSIMISTIC: pessimism, sadness OTHER: anticipation, joy, love, optimism, surprise, trust

Table 3: Mapped attitudinal function categories in the emotion detection datasets.

3.3 Experimental setup: supervised approaches

The experiment was conducted employing several supervised approaches, including traditional machine learning, fine-tuned Transformers, negative attitudinal function metaembeddings, and zero-shot classification through Natural Language Inference (NLI). Baseline models were developed, using these approaches, with the train splits of each mapped emotion detection datasets. Negative attitudinal function models were developed, using these approaches, with the synthetic dataset. The evaluation was conducted with the test splits of each mapped emotion detection dataset. A brief explanation is given for each approach:

- **Traditional Machine Learning with Multinomial Naïve Bayes (NB)**: We trained a Multinomial Naïve Bayes model using bag-of-words features for text classification tasks.
- **Fine-tuning Transformers (T-XLM-R)**: We utilized XLM-T (Barbieri, Anke, and Camacho-Collados,

2022), a variant of XLM-RoBERTa (Conneau et al., 2020) pre-trained with tweets in over 30 languages. For single-label tasks, we fine-tuned the model with a softmax classifier, while for multi-label tasks, we used binary cross-entropy loss or a softmax classifier with an adapted loss function.

- **Negative attitudinal function metaembeddings**: Contextualized embeddings were obtained using Paraphrase Multilingual Mpnet Base v2 for semantic similarity tasks (Reimers and Gurevych, 2020). Mean pooling was applied to compute attitudinal function metaembeddings for each negative attitude function category on the basis of the samples of the synthetic dataset, and cosine similarity determined the semantic similarity between these metaembeddings and the sentence embeddings obtained from the tweets.

- **Zero-shot classification with NLI**: We employed mDeBERTa v3 fine-tuned on NLI tasks for zero-shot classification (He, Gao, and Chen, 2021). Prompt engineering involved 13 prompts focusing on emotional states, text emotions, functions, or intentions expressed in the text (Table 4). The candidate labels were the emotion categories, then mapped to functions.

Prompts
This person feels _.
This person conveys _.
This person shows _.
This person expresses _.
This text is _.
This text is about _.
This text shows _.
This text expresses _.
This text conveys _.
The communicative function of this text is _.
The communicative intention of this text is _.
The emotion of this text is _.
The emotion expressed in this text is _.

Table 4: Prompts used in our zero-shot classification approach with NLI.

4 Results and discussion

Table 5 offers the results of each experimental setup for each dataset. In bold, the best scores achieved for each dataset are highlighted.

Baseline models often provided the best results due to the use of in-domain data from emotion detection datasets. On the other hand, the function models were expected to behave worse, due to their evaluation with out-of-domain data and the qualitative differences between the tasks of emotion detection and negative attitudinal identification. Despite that, we highlight the robust performance of the metaembedding function model with the CARER and AIT datasets, as they are loaded with many explicit emotion expressions, approaching the fine-tuned baseline. However, the function models displayed subpar performance with EmoEvent and GoEmotions, possibly due to dataset annotation issues impacting performance, as seen in the scores obtained by the baseline models and our error analysis stage. The best-performing prompt in the zero-shot NLI approach was *The emotion expressed in this text is* {*label*}., except for GoEmotions, which was *This text is* {*label*}.

Notably, our function models, developed with the negative attitudinal constructions in English, showcased remarkable generalization to tweets in Spanish and Arabic, as shown by the results obtained in both the fine-tuning and metaembedding approaches of the function models in the AIT dataset. This may suggest the universal applicability of emotional expressions encoded in some function constructions, potentially obviating the need to build function datasets in other languages.

In summary, our models using the synthetic dataset of negative attitudinal functions exhibit promising capabilities in emotion detection for unseen data and multilingual contexts.

4.1 Limitations, challenges, and future research directions

Our evaluation of function datasets against emotion detection baselines faced challenges due to the qualitative differences of the tasks and annotation issues found in some emotion detection datasets. Mapping emotions to functions was not consistently equivalent, impacting fairness. Annotation issues in the EmoEvent and GoEmotions datasets affected the performance of all models. CARER and AIT, with higher quality annotations, led to

Dataset	Model	Evaluation metrics		
		Macro-F1	Micro-F1	Weighted-F1
EmoEvent en-es	NB-baseline	0.25	0.70	0.66
	NB-basic	0.16	0.38	0.46
	T-XLM-R-baseline	**0.38**	0.70	**0.69**
	T-XLM-R-basic	0.18	**0.74**	0.65
	Metaembedding-baseline	0.27	0.49	0.56
	Metaembedding-basic	0.25	0.56	0.59
	Zero-shot NLI	0.26	0.45	0.52
GoEmotions	NB-baseline	0.12	0.75	0.64
	NB-basic	0.12	0.38	0.42
	T-XLM-R-baseline	**0.52**	**0.78**	**0.78**
	T-XLM-R-basic	0.18	0.28	0.66
	Metaembedding-baseline	0.21	0.31	0.53
	Metaembedding-basic	0.23	0.36	0.49
	Zero-shot NLI	0.23	0.15	0.09
CARER	NB-baseline	0.77	0.84	0.82
	NB-basic	0.26	0.38	0.34
	T-XLM-R-baseline	**0.95**	**0.96**	**0.96**
	T-XLM-R-basic	0.23	0.48	0.33
	Metaembedding-baseline	0.62	0.67	0.67
	Metaembedding-basic	0.45	0.52	0.51
	Zero-shot NLI	0.58	0.64	0.65
AIT en-es-ar	NB-baseline	0.53	0.60	0.58
	NB-basic	0.16	0.19	0.18
	T-XLM-R-baseline	**0.75**	**0.77**	**0.77**
	T-XLM-R-basic	0.43	0.49	0.48
	Metaembedding-baseline	0.48	0.49	0.53
	Metaembedding-basic	0.50	0.54	0.54
	Zero-shot NLI	0.54	0.55	0.54

Table 5: Results of the experiments.

better model performance.

Improving emotion dataset annotation quality through manual supervision could enhance model performance. Another line of research could focus on manually tagging the emotion detection datasets with functions, using expert annotation and/or crowdsourcing. Future research could also focus on synthetic data generation (Dai et al., 2023) using large language models (LLMs) like ChatGPT or BARD, which can also be used as automatic annotators (Ostyakova et al., 2023b; Kaddour et al., 2023). We could also build tweet datasets with semi-automatic methods by leveraging the lexico-grammatical patterns found in our taxonomy and use expert annotation for revising the samples obtained. Human annotated and synthetically generated data could be compared and combined in future experiments. Future work could also employ semantic role labeling with frames (Fillmore, 2006; Baker, Fillmore, and Lowe, 1998) to enhance attitudinal function conceptualization and granularity. Another line of research could focus on expanding and/or creating new function taxonomies for diverse purposes, contexts, and languages.

4.2 Ethical considerations

The increasing use of affect-related data by governments and corporations raises ethical concerns, particularly in the context of digital surveillance. Users willingly disclose extensive personal information on social media (Han, 2015), leading to potential misuse, such as predicting and manipulating users' behavior for advertising and political purposes (McStay, 2020). This 'psychological targeting' involves building psychological profiles from digital footprints for manipulative ends (Matz, Appel, and Kosinski, 2020). The ethical implications include the infringement of freedom and harm to individuals' interests. Proposed solutions involve limiting exposure, raising public awareness, and enforcing regulations (McStay, 2020). Ethical decision-making for affect-related systems should consider task design, data building, and annotation processes (Mohammad, 2022). The synthetic dataset was created to avoid privacy issues concerning the collection of user-generated data. Access to our dataset should prioritize social good and ethical considerations. Our negative attitudinal function identification task aims to enhance users' well-being in smart city contexts, not manipulate users' behavior.

5 Conclusion

We introduce the task of negative attitudinal function identification, which seeks to discern users' emotional reactions, judgments, and appreciations. It holds potential for smart city scenarios, where citizens' expressions of emotions, judgements, and appreciations can inform policies for social good, addressing concerns and enhancing well-being. Negative attitudinal functions enable citizens to voice complaints about various issues. Unlike emotion detection and dialog act classification, this task draws on resources from and insights into theoretical and applied linguistic research, offering a more nuanced understanding of attitudinal intent. Beyond blending the interests of both emotion detection and dialog act classification, our task widens its scope to include subjective states that do not necessarily carry emotional meaning, such as judgments and appreciations.

We constructed a linguistically informed synthetic dataset of negative attitudinal functions that contained lexico-grammatical patterns. This dataset was then used to develop different supervised approaches. Emotion detection datasets of tweets were reused for our task by mapping their emotion categories to the functions of our taxonomy. They were then used to develop different supervised approaches. This was the baseline. An evaluation stage was conducted to compare the performance of the function models against the baseline. Results revealed promising capabilities of our function models, despite challenges such as qualitative differences between the tasks, the use of out-of-domain data, and annotation noise. Despite these challenges, negative attitudinal function models demonstrate promising potential, particularly in multilingual contexts, and with out-of-domain data. We hope that NLP practitioners and researchers can benefit from this new NLP task and the associated synthetic dataset of negative attitudinal functions.

Acknowledgements

Financial support for this research has been provided under grant PID2020-112827GB-I00, funded by the Spanish Ministry of Science and Innovation MCIN/AEI/10.13039/501100011033, and grant number 101017861[project SMART-LAGOON], funded by the European Union's Horizon 2020 research and innovation program.

References

Alba-Juez, L. 2018. Emotion and appraisal processes in language. In M. d. l. Á. Gómez González and J. L. Mackenzie, editors, *The Construction of Discourse as Verbal Interaction*. John Benjamins, pages 227–250.

Alm, C. O., D. Roth, and R. Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *HLT/EMNLP 2005 - Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, number October, pages 579–586.

Aman, S. and S. Szpakowicz. 2007. Identifying expressions of emotion in text. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4629 LNAI, pages 196–205. Springer Berlin Heidelberg.

Aroyehun, S. T., L. Malik, H. Metzler, N. Haimerl, A. Di Natale, and D. Garcia. 2023. LEIA: Linguistic Embeddings for the Identification of Affect. *CoRR*, cs.CL/2304:1–15, apr.

Baker, C. F., C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet Project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Baldwin, T., P. Cook, M. Lui, A. MacKinlay, and L. Wang. 2013. How Noisy Social Media Text, How Diffrnt Social Media Sources? *International Joint Conference on Natural Language Processing*, (October):356–364.

Barbieri, F., L. E. Anke, and J. Camacho-Collados. 2022. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *2022 Language Resources and Evaluation Conference, LREC 2022*, number June, pages 258–266, Marseille, France. European Language Resources Association.

Basile, A., G. Pérez-Torró, and M. Franco-Salvador. 2021. Probabilistic Ensembles of Zero- and Few-Shot Learning Models for Emotion Classification. *International Conference Recent Advances in Natural Language Processing, RANLP*, pages 128–137.

Bednarek, M. 2008. *Emotion Talk across Corpora*. Palgrave Macmillan.

Blundell, J., J. Higgens, and N. Middlemiss. 1982. *Function in English*. Oxford University Press.

Bostan, L., E. Kim, and R. Klinger. 2020. Good news everyone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, number May, pages 1554–1566, Marseille, France. European Language Resources Association.

Bostan, L.-A.-M. and R. Klinger. 2018. An Analysis of Annotated Corpora for Emotion Classification in Text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119.

Buechel, S. and U. Hahn. 2017. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. In *LAW 2017 - 11th Linguistic Annotation Workshop, Proceedings of the Workshop*, pages 1–12. Association for Computational Linguistics.

Bunt, H. 2009. The DIT++ taxonomy for functional dialogue markup. *Proceedings of EDAML/AAMAS Workshop "Towards a Standard Markup Language for Embodied Dialogue Acts"*, pages 13–24.

Bunt, H. 2019. Guidelines for using ISO standard 24617-2. Technical report.

Bunt, H., V. Petukhova, A. Malchanau, A. Fang, and K. Wijnhoven. 2016. The DialogBank. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 3151–3158, Portorož, Slovenia. European Language Resources Association (ELRA).

Cambria, E. 2016. Affective Computing and Sentiment Analysis. *IEEE Intelligent Systems*, 31(2):102–107.

Campagnano, C., S. Conia, and R. Navigli. 2022. SRL4E – Semantic Role Labeling for Emotions: A Unified Evaluation Framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 4586–4601, Dublin, Ireland. Association for Computational Linguistics.

Cerisara, C., S. Jafaritazehjani, A. Oluokun, and H. Le. 2018. Multi-task dialog act and sentiment recognition on Mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 745–754, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Core, M. G. and J. F. Allen. 1997. Coding Dialogs with the DAMSL Annotation Scheme. In *Proceedings of the 1997 AAAI Fall Symposium: Communicative Action in Humans and Machines*, pages 28–35.

Dai, H., Z. Liu, W. Liao, X. Huang, Z. Wu, L. Zhao, W. Liu, N. Liu, S. Li, D. Zhu, H. Cai, Q. Li, D. Shen, T. Liu, and X. Li. 2023. ChatAug: Leveraging ChatGPT for Text Data Augmentation. pages 1–12.

Demszky, D., D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi. 2020. GoEmotions: A Dataset of Fine-Grained

Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, pages 4171–4186.

Dini, L. and A. Bittar. 2016. Emotion analysis on twitter: The hidden challenge. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 3953–3958, Portorož, Slovenia. European Language Resources Association (ELRA).

Dwivedi-Yu, J. and A. Y. Halevy. 2022. "That's so cute!": The CARE Dataset for Affective Response Detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 50–69. Association for Computational Linguistics.

Ekman, P. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, may.

Enayet, A. and G. Sukthankar. 2020. A Transfer Learning Approach for Dialogue Act Classification of GitHub Issue Comments.

Fillmore, C. J. 2006. Frame semantics. In D. Geeraerts, editor, *Cognitive linguistics: Basic readings*, volume 34. Mouton de Gruyter, Berlin, New York, pages 373–400.

Finocchiaro, M. and C. Brumfit. 1983. *The functional-notional approach: from theory to practice*. Oxford University Press, New York.

Guiraud, N., D. Longin, E. Lorini, S. Pesty, and J. Rivière. 2011. The face of emotions: A logical formalization of expressive speech acts. In Y. Tumer, P. Yolum, L. Sonenberg, and P. Stone, editors, *10th International Conference on Autonomous Agents and Multiagent Systems 2011 (AAMAS 2011)*, volume 2, pages 969–976, Taipei, Taiwan. International Foundation for Autonomous Agents and Multiagent Systems.

Gung, J., R. Shu, E. Moeng, W. Rose, S. Romeo, A. Gupta, S. Mansour, and Y. Zhang. 2023. Intent Induction from Conversations for Task-Oriented Dialogue at DSTC 11. *CoRR*, cs.CL/2304.

Han, B.-C. 2015. *The Transparency Society*. Stanford University Press, Stanford.

He, P., J. Gao, and W. Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *CoRR*, cs.CL/2111:1–17.

Jurafsky, D. and J. H. Martin. 2023. Chatbots Dialogue Systems. In *Speech and Language Processing*. Book in preparation, chapter 15.

Jurafsky, D., E. Shriberg, and D. Biasca. 1997. Switchboard SWBD DAMSL Shallow-Discourse-Function Annotation. Technical report.

Jurafsky, D., E. Shriberg, B. Fox, and T. Curl. 1998. Lexical, prosodic, and syntactic cues for dialog acts. *ACL/COLING Workshop on Discourse Relations and Discourse Markers*, pages 114–120.

Kaddour, J., J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy. 2023. Challenges and Applications of Large Language Models. (1).

Khanpour, H., N. Guntakandla, and R. Nielsen. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, pages 2012–2021.

Leech, G. and M. Weisser. 2003. Generic Speech Act Annotation for Task-Oriented Dialogue. In D. Archer, P. Rayson, A. Wilson, and A. McEnery, editors, *Proceedings of the Corpus Linguistics 2003 Conference*, pages 441–446, Lancaster. UCREL Technical Papers.

Li, Y., H. Su, X. Shen, W. Li, Z. Cao, and S. Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the The 8th International*

Joint Conference on Natural Language Processing, pages 986–995, Taipei, Taiwan. AFNLP.

Liew, J. S. Y. and H. R. Turtle. 2016. Exploring Fine-Grained Emotion Detection in Tweets. In *Proceedings of NAACL-HLT*, pages 73–80, San Diego, California. Association for Computational Linguistics.

Liew, J. S. Y., H. R. Turtle, and E. D. Liddy. 2016. EmoTweet-28 : A Fine-Grained Emotion Corpus for Sentiment Analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1149–1156, Portorož, Slovenia. European Language Resources Association (ELRA).

Maia, B. and D. Santos. 2018. Language, emotion, and the emotions: The multidisciplinary and linguistic background. *Language and Linguistics Compass*, 12(6):1–17.

Martin, J. R. and P. R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan.

Matz, S. C., R. E. Appel, and M. Kosinski. 2020. Privacy in the age of psychological targeting. *Current Opinion in Psychology*, 31:116–121.

McStay, A. 2020. Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. *Big Data and Society*, 7(1).

Milanovic, M. and N. Saville. 2012. The Theoretical Foundations for functions in the Council of Europe modern languages projects and the Common European Framework of Reference for languages. In *Language Functions Revisited*. Cambridge University Press.

Mohammad, S. M. 2012. Emotional tweets. In **SEM 2012 - 1st Joint Conference on Lexical and Computational Semantics*, volume 1, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Mohammad, S. M. 2021. Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. In *Emotion Measurement*. Elsevier, chapter 11, pages 201–237.

Mohammad, S. M. 2022. Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis. *Computational Linguistics*, 48(2):239–278.

Mohammad, S. M. and C. O. Alm. 2015. Computational Analysis of Affect and Emotion in Language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, Lisbon, Portugal. Association for Computational Linguistics.

Mohammad, S. M., F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings ofthe 12th International Workshop on Semantic Evaluation (SemEval-2018)*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Mohammad, S. M. and S. Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 198–209.

Mohammad, S. M., X. Zhu, S. Kiritchenko, and J. Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing and Management*, 51(4):480–499.

Ostyakova, L., K. Petukhova, V. Smilga, and D. Zharikova. 2023a. Linguistic Annotation Generation with ChatGPT: a Synthetic Dataset of Speech Functions for Discourse Annotation of Casual Conversations ChatGPT : . In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2023"*, pages 1–18.

Ostyakova, L., V. Smilga, K. Petukhova, M. Molchanova, and D. Kornev. 2023b. ChatGPT vs. Crowdsourcing vs. Experts : Annotating Open-Domain Conversations with Speech Functions. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 242–254. Association for Computational Linguistics.

Periñán-Pascual, C. 2023. From Smart City to Smart Society: A quality-of-life ontological model for problem detection from user-generated content. *Applied Ontology*, pages 1–44.

Picard, R. W. 1997. *Affective computing.* MIT Press, Cambridge, MA, US.

Plaza del Arco, F. M., M.-T. Martín-Valdivia, and R. Klinger. 2022. Natural Language Inference Prompts for Zero-shot Emotion Classification in Text across Corpora. *CoRR*, cs.CL/2209.

Plaza del Arco, F. M., C. Strapparava, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. 2020. EmoEvent: A multilingual emotion corpus based on different events. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, number May, pages 1492–1498, Marseille, France. European Language Resources Association.

Plutchik, R. 1980. A general psychoevolutionary theory of emotion. In R. Plutchik and H. B. T. T. o. E. Kellerman, editors, *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion.* Academic Press, chapter A GENERAL, pages 3–33.

Purver, M. and S. Battersby. 2012. Experimenting with distant supervision for emotion classification. In *EACL 2012 - 13th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, pages 482–491, Avignon, France. Association for Computational Linguistics.

Quan, C. and F. Ren. 2009. Construction of a blog emotion corpus for Chinese emotional expression analysis. In *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009*, number August, pages 1446–1454.

Reimers, N. and I. Gurevych. 2020. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 3982–3992.

Roberts, K., M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu. 2012. EmpaTweet: Annotating and detecting emotions on twitter. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pages 3806–3813, Istanbul, Turkey. European Language Resources Association (ELRA).

Ruby, D. 2023. 58+ Twitter Statistics For Marketers In 2023 (Users Trends).

Saha, T., S. Ramesh Jayashree, S. Saha, and P. Bhattacharyya. 2020. BERT-Caps: A Transformer-Based Capsule Network for Tweet Act Classification. *IEEE Transactions on Computational Social Systems*, 7(5):1168–1179.

Saravia, E., H. C. Toby Liu, Y. H. Huang, J. Wu, and Y. S. Chen. 2018. Carer: Contextualized affect representations for emotion recognition. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 3687–3697.

Sbisà, M. 1975. Speech acts and the expression of affect.

Scherer, K. R. 2005. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, dec.

Scherer, K. R. and H. G. Wallbott. 1994. Evidence for universality and cultural variation. *Journal of Personality and Social Psychology*, 66(2):310–328.

Schuff, H., J. Barnes, J. Mohme, S. Padó, and R. Klinger. 2017. Annotation, Modelling and Analysis of Fine-Grained Emotions on a Stance and Sentiment Detection Corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 13–23, Copenhagen, Denmark. Association for Computational Linguistics.

Searle, J. R. 1976. A classification of illocutionary Acts. *Language in Society*, 5(1):1–24.

Semeraro, A., M. Stella, A. Semeraro, S. Vilella, S. M. Mohammad, G. Ruffo, and M. Stella. 2023. EmoAtlas: An emotional profiling tool merging psychological lexicons, artificial intelligence and network science.

Sorabji, R. 2002. *Emotion and Peace of Mind: From Stoic Agitation to Christian Temptation*. Oxford University Press.

Stolcke, A., K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):35.

Strapparava, C. and R. Mihalcea. 2007. Learning to identify emotions in text. In *Proceedings of the ACM Symposium on Applied Computing*, pages 1556–1560.

Troiano, E., L. Oberländer, and R. Klinger. 2023. Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction. *Computational Linguistics*, 49(1):1–72.

Vera, D., O. Araque, and C. A. Iglesias. 2021. GSI-UPM at IberLEF2021: Emotion analysis of Spanish tweets by fine-tuning the XLM-RoBERTa language model. In *CEUR Workshop Proceedings*, volume 2943, pages 16–26.

Vosoughi, S. and D. Roy. 2016. Tweet acts: A speech act classifier for Twitter. In *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, pages 711–714.

Wang, W., L. Chen, K. Thirunarayan, and A. P. Sheth. 2012. Harnessing twitter 'big data' for automatic emotion identification. In *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, pages 587–592.

Wegge, M., E. Troiano, L. Oberländer, and R. Klinger. 2022. Experiencer-Specific Emotion and Appraisal Prediction. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 25–32. Association for Computational Linguistics.

Weisser, M. 2018. *How to Do Corpus Pragmatics on Pragmatically Annotated Data*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Wilkins, D. A. 1976. *Notional Syllabuses*. Oxford University Press.

Yang, K., S. Ji, T. Zhang, Q. Xie, and S. Ananiadou. 2023. On the Evaluations of ChatGPT and Emotion-enhanced Prompting for Mental Health Analysis. *CoRR*, cs.CL/2304.

Yu, D., L. Li, H. Su, and M. Fuoli. 2023. Using LLM-assisted Annotation for Corpus Linguistics A Case Study of Local Grammar Analysis. *CoRR*, cs.CL/2305.

Yu, D. and Z. Yu. 2021. MIDAS: A dialog act annotation scheme for open domain human machine spoken conversations. In *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 1103–1120. Association for Computational Linguistics.

Żelasko, P., R. Pappagari, and N. Dehak. 2021. What helps transformers recognize conversational structure? Importance of context, punctuation, and labels in dialog act recognition. *Transactions of the Association for Computational Linguistics*, 9:1179–1195.

Zhabotynska, S. and N. Slyvka. 2020. Emotive speech acts and their discourse modifications in the literary text. *Discourse and Interaction*, 13(1):113–136.

Zhang, R., D. Gao, and W. Li. 2011. What are tweeters doing: Recognizing speech acts in twitter. In *AAAI Workshop - Technical Report*, volume WS-11-05, pages 86–91.

Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 73-85 recibido 14-12-2023 revisado 10-01-2024 aceptado 28-01-2024

Gradable ChatGPT Translation Evaluation

Evaluación de traducción graduada de ChatGPT

Hui Jiao,[1] Bei Peng,[2] Lu Zong,[1] Xiaojun Zhang,[1,4] Xinwei Li[3]
[1]Xi'an Jiaotong-Liverpool University
[2]University of Liverpool
[3]Southeast University
[4]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
Xiaojun.Zhang01@xjtlu.edu.cn

Abstract: ChatGPT, as a language model based on large-scale pre-training, has exerted a profound influence on the domain of machine translation. In ChatGPT, a "Prompt" refers to a segment of text or instruction employed to steer the model towards generating a specific category of response. The design of the translation prompt emerges as a key aspect that can wield influence over factors such as the style, precision and accuracy of the translation to a certain extent. However, there is a lack of a common standard and methodology on how to design and select a translation prompt. Accordingly, this paper proposes a generic taxonomy, which defines gradable translation prompts in terms of expression type, translation style, Part-of-Speech information and explicit statement, thus facilitating the construction of prompts endowed with distinct attributes tailored for various translation tasks. Specific experiments and cases are selected to validate and illustrate the effectiveness of the method.
Keywords: Translation prompt, ChatGPT, T3S taxonomy, Evaluation.

Resumen: ChatGPT, un modelo de lenguaje basado en un pre-entrenamiento a gran escala que ha tenido un profundo impacto en la traducción automática. En este contexto, un "prompt" se refiere a un segmento de texto o instrucción utilizada para dirigir el modelo hacia la generación de una respuesta específica. El diseño del prompt de traducción es crucial y puede influir en aspectos como el estilo, la precisión y la exactitud de la traducción. Sin embargo, actualmente carecemos de un estándar y metodología común para diseñar y seleccionar prompts de traducción. Por lo tanto, este artículo propone una taxonomía genérica que define prompts de traducción evaluables en términos de expresión, estilo, Part-Of-Speech (POS) y declaración explícita, facilitando la construcción de prompts con distintos atributos adaptados a diversas tareas de traducción. Se han seleccionado experimentos y casos específicos para validar e ilustrar la eficacia del método.
Palabras clave: Prompt de traducción, ChatGPT, Taxonomía T3S, Evaluación.

1 Introduction

Machine translation (MT), one of the oldest branches of research in the field of natural language processing, involves techniques for transforming one natural language into another. As a key research area within the field of Artificial Intelligence, MT has been widely used in a wide range of fields and has attracted extensive attention from both academia and industry (Yang, Wang, and Chu, 2020). In recent years, there has been a growing trend towards the use of large-scale pretrained language models for natural language processing (NLP) (Yang, Wang, and Chu, 2020; Brown et al., 2020; Amplayo, Yoo, and Lee, 2022). Large Language Models (LLMs), such as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), and LLaMA (Tou-

Corresponding author: Xiaojun Zhang

ISSN 1135-5948 DOI 10.26342/2024-72-5

abundant representations of the input text, and have greatly contributed to the development of MT technology (Lewis et al., 2019).

NLP tasks have recently been significantly influenced by the emergence of ChatGPT, a powerful pre-trained LLM developed by OpenAI. The model has been trained to perform a large number of human-like tasks (e.g., question answering, code debugging, generating evaluations, etc.) with human feedback. However, the overall performance of ChatGPT on translation tasks using simple prompts and basic settings is not as good as the commercial translation products such as Google Translate and Microsoft Translate (He et al., 2022; Zhou et al., 2022; Jiao et al., 2023; Hendy et al., 2023)but is promising to surpass them with complex and explicit prompts. This is due to the fact that for a given complex translation task, significant differences in the performance of LLMs arise when different types/styles/levels of detail of prompts are introduced (Karmaker and Feng, 2023).

The effective use of LLMs requires elaborate prompt engineering, which refers to the process of designing and refining the prompts or instructions provided to a large language model (Zhou et al., 2022; Liu et al., 2023). For translation tasks of different styles, the relevant prompts that the user can try while performing the task are different. In addition, the number of details included in the prompts also largely affects the performance of LLMs when dealing with target complex translation tasks (Karmaker and Feng, 2023). By providing clear and well-structured prompts, users can help guide LLMs in the right direction and reduce potential biases or errors. To achieve this, we propose a unified gradable prompting taxonomy for ChatGPT translation called T3S, which employs standardized criteria to categorize various types of translation prompts, thereby further enhancing ChatGPT's translation capabilities.

On the one hand, ChatGPT is utilized to conduct very broad and various types of translation tasks from general text translation to highly specialised domains (e.g., medical, legal, technical, etc.), and categorising prompts will help ChatGPT distinguish between different prompts and implement different translation strategies to meet translation needs in different domains and contexts. On the other hand, the proposed T3S taxonomy can help researchers explore in depth the working principles of LLMs and their performance differences in different domains, thus promoting further research and innovation in LLMs. Specifically, the widespread adoption of this taxonomy can potentially promote a more accurate performance assessment of ChatGPT in different translation tasks, thereby identifying specific problems that may exist in the model, providing targeted feedback for improvement and guiding the direction of model optimisation, and achieving continuous progress of the model.

2 Related Work

A prompt is a set of instructions provided to an LLM that enhances the functionality of the LLM by customising it (Liu et al., 2023). In the field of large language modelling, complex tasks refer to those involving multiple steps or subtasks that require a higher level of semantic understanding, planning, reasoning, and natural language generation capabilities, which makes prompt engineering particularly critical and challenging (Tan et al., 2022).

In recent years, many researchers have proposed different approaches to engineering prompts. For example, one of the best-known (and easiest to implement) prompt engineering techniques is to add "Think step by step" to the end of the prompt. Adding this phrase improves the accuracy of the GPT-3 (text-DaVinci-002 model) across multiple tasks (Wei et al., 2022). Moreover, Brown et al. (2020) presented a standard question-answer pair prompting technique which produces a few-shot effect. By providing suitable output instances, LLMs are more likely to produce the desired output (Zhao et al., 2021). Similar to "Think step by step", the Chain of Thought (CoT) prompting method guides LLMs to break down a complex task into multiple intermediate steps (Wang et al., 2023). Fu et al. (2022)showed that separating each step with a new line in exemplar reasoning is much more effective than separating each step with a full stop. Researchers have also explored other prompt design techniques such as Reasoning and Acting (ReAct) (Yao et al., 2022), which overcomes the illusions and error propagation problems prevalent in CoT reasoning by interacting with a simple Wikipedia API. Other techniques such as Zero-shot-CoT (Kojima et al., 2022) and Self-Ask (Press et al., 2022), improve

LLMs' reasoning and action in solving questions and answering tasks. Meanwhile, Kim, Baldi, and McAleer (2023) suggested that recursively criticising and improving its output (RCI) is superior to CoT prompts in terms of its effectiveness in reasoning ability in a range of natural language reasoning tasks.

More specifically, conversational LLMs, such as ChatGPT, have generated considerable research interest in a range of domains, with tasks ranging from answering questions for medical licensing exams to generating code snippets (Gilson et al., 2023). Correct prompt engineering has become a key skill for users wishing to utilise the full potential of ChatGPT and obtain optimal results in a variety of applications. Accordingly, there has been an influx of prompt engineering research in many different areas. Thirunavukarasu et al. (2023) explored how prompts can be used to enhance the efficiency and effectiveness of ChatGPT in medical clinics, education, and research, while Trautmann, Petrova, and Schilder (2022) proposed zero-sample legal prompts engineering (LPE) to guide and enhance LLMs in natural legal language processing (NLLP) capabilities. These studies focused on NLP tasks rather than MT. In the field of machine translation, however, previous studies (Liu et al., 2019; Guo et al., 2020) have shown that while LLMs can enhance a translation system's understanding of the source text, improving its generative capabilities is more difficult. A well-developed translation system requires strong language comprehension and generation capabilities to achieve accurate and fluent translation results. Although some studies (Brown et al., 2020; Chowdhery et al., 2022) have explored the effects of different prompts on translation results, there is still a lack of systematic research on how to improve MT using prompts. With the popularity of LLM-based prompting approaches, researchers are starting to recognise the importance of introducing prompts into neural machine translation (NMT) (Li et al., 2022; Tan et al., 2022; Wei et al., 2022). Nevertheless, these approaches still rely on pre-training or fine-tuning the models, rather than directly applying them to "frozen" LLMs. Therefore, it is critical to study how to make the most effective use of these prompts in order to balance language comprehension and generation capabilities and achieve better results in MT.

To summarise, prompt engineering is a crucial step for the effective utilisation of LLMs. However, existing works have mainly emphasised the use of diversified prompts to improve the ability of LLMs to perform general-purpose NLP tasks, while the specification of prompts for machine translation remains under-explored. Therefore, this paper aims to fill the research gap by providing translators with a systematic approach to selecting and designing prompts, which can improve the consistency, reliability and quality of ChatGPT translations, as well as promote the development and innovation in the field of MT.

3 ChatGPT Translation Prompting Taxonomy Design

Through training, ChatGPT is capable of generating appropriate responses based on the given prompts. This attribute determines its high sensitivity to the information provided by the prompts. Differences in several key factors of the prompts can have a significant impact on the accuracy and performance of large language models in translation tasks (Karmaker and Feng, 2023). These key factors are described below, where the prompt is defined by the combination of the instruction (intended target) and the source text provided to ChatGPT for performing the translation task. The prompts present differences in the instruction section under the assumption that the source text has been provided.

Ambiguous or contextually inadequate prompts can easily confuse LLMs, leading to inaccurate or irrelevant responses from these models (Jiang et al., 2022). Therefore, providing clear and specific instructions in the prompts can help guide ChatGPT to generate the desired translation results. In the MT domain, detailed prompts usually consist of elaborating on various aspects of the translation task specification. In the following, we will delve into two basic aspects of translation task specification.

1) Explicit Descriptions. In translation task prompts, a clear task description is essential for obtaining accurate and relevant translation results. Specifying the translation objectives clearly and asking ChatGPT to proofread before generating the translation results helps to guide ChatGPT to a full understanding of the translation task and maintain a consistent focus throughout the translation process. This ensures a higher level of accuracy, relevance and fluency in the translation results, thereby increasing the probability of obtaining the desired translation outcome.

2) Contextual Information. Contextual information plays a key role in enhancing ChatGPT's understanding of words, phrases, and sentences in the source text, thus helping to reduce ambiguity and misunderstanding. Models can optimise translation decisions based on context, so as to avoid translation errors (Popescu-Belis, 2019). For example, for polysemous words, context can help ChatGPT determine the correct word meaning. This is crucial for producing accurate translations, as the same word may be translated differently in different contexts. Additionally, contextual background helps to maintain consistency in the model's expression throughout the translation process. When the source text has certain specific usages or terms in the context, the model can retain them in the translation based on the context, ensuring a coherent translation.

Translation prompts consist of instructions and source text. Assuming that the source text is invariant in a given translation task, we believe that the difference between prompts lies in the explicit descriptions and contextual information they contain. To address the key factors of prompt selection and design, we will classify the prompts for ChatGPT translation tasks according to the following four aspects.

3.1 Expression Types

There are two main types of expressing translation prompts: the single-turn prompt and the multiple-turn prompt. Both types of expression can be effective in different contexts. Single-turn prompts involve presenting the model with a solitary input, typically in the form of a single sentence or a brief textual segment ("*Please translate the following text...*"), for the model to translate into the target language. Whereas multiple-turn prompts incorporate conversational interactions, usually consisting of multiple dialogue rounds, so that the model can better understand and perform the translation task. For example, a user can ask ChatGPT to check and revise the translation after it responds to the first round of prompts ("*Please translate it again / Please revise the translation*"). Such prompts can be used for more complex translation tasks, where context, clarification, follow-up questions, etc., may need to be taken into account. Compared to single-turn prompts, the use of multiple-turn prompts can significantly enhance the comprehension of ChatGPT, and reduce its tendency to generate irrelevant or inaccurate responses (Pan et al., 2023). Depending on the nature and needs of the translation task, choosing the appropriate prompt expression form can help ChatGPT generate the required translation output more accurately.

3.2 Translation Style

Depending on the translation task in different translation domains to deal with different genres such as literary, medicine, legal and commercial texts, the desired prompt needs to be defined and selected according to the translation style. Defining translation style involves determining the content and expression to be conveyed in a translation task. It relates to the specific translation target and audience. Encompassing facets such as affective undertones, tonal modulations, and the amplitude of linguistic exposition, the election of a fitting translation style exerts a discernible impact on the ultimate rendition. Thus, the inclusion of relevant contextual information such as target audience and level of expression in the prompt can provide ChatGPT with additional information to produce a more accurate translation result.

3.3 POS Information

Part of Speech (POS) is a grammatical category that includes nouns, verbs, adjectives, adverbs, etc. (Hlaing et al., 2022). Many NLP tasks benefit from the use of POS tags. In translation tasks, POS tags help ChatGPT to capture the grammatical structure of a sentence in the source language, and accurately locate the grammatical roles of each word in the sentence, so as to better translate it into the grammatical structure of the target language, eliminating word ambiguities, and further enhancing the natural fluency of the translation result. For instance, Feng et al. (2020) demonstrated that incorporating POS tagging information into the target side can significantly improve the translation performance of the NMT system in both Chinese-to-English and German-to-English translation pairs. Further, Hlaing et al. (2022) conducted an NMT study using POS tagging information on low-resource language pairs, explicitly pointing out the necessity of integrating POS tags when using NMT models that include linguistic features.

3.4 Few-shot Prompts

LLMs can benefit from example-based learning, which involves providing specific input-output pair examples (a small number of examples). This can help models to better understand task requirements and generate appropriate output (Brown et al., 2020). We believe that including a few input-output examples in the prompt will improve the performance of the LLMs without any adjustments to the parameters or architecture. For example, by including specific terminology and styles in the few-shot prompts, ChatGPT is able to adapt quickly across different translation domains, generating domain-appropriate translations.

We propose a gradable prompting taxonomy for ChatGPT translation which is categorised into five different levels based on the above four key elements in prompt design including expression type, translation style, POS information, and few-shot prompts. We named it T3S standing for expression type, translation style, POS information and few-shot 5-level prompting taxonomy.

More precisely, level "0" represents the lowest level of detail, where only basic translation is required with the general prompt such as "*Please translate the following text...*"; Level "1" distinguishes the expression type of the single-turn form with the multiple-turn form; Level "2" adds translation style instruction providing contextual information; Level "3" integrates POS information into translation instruction; and Level "4" represents the highest level of detail, where the multiple-turn prompts include clear instructions, explicit contextual information of translation with few-shot examples, and an explicit statement asking ChatGPT to check and revise the results.

4 Experimental Validation

In order to verify the rationality and validity of the T3S taxonomy, we set up an evaluation experiment. In the following, we show the details of the experimental setup, including the adopted dataset and evaluation metrics. The results and analyses of the experiment are also presented.

4.1 Dataset

We evaluated the translation quality of ChatGPT at different levels of prompt on the Flores-101 (Goyal et al., 2022)dataset. The dataset consists of 1012 sentences extracted from the English Wikipedia covering a wide variety of topics and domains. In real translation applications, ChatGPT needs to process texts from a variety of topics and domains. A dataset covering different topic domains can help us evaluate ChatGPT's generalisation ability and gain a more comprehensive understanding of ChatGPT's translation performance under different contexts. Moreover, these sentences have been translated into 101 languages by professional translators through a rigorously controlled process with automated and manual quality checks. Furthermore, all translations are multilingual aligned. Such a high-quality and high-coverage dataset ensures the accuracy and consistency of reference translations, and better helps us understand and evaluate the quality and performance of ChatGPT's translations. However, a graded translation quality assessment for multiple languages may add complexity and resource requirements. Therefore, for a clearer direct comparison of translation quality and to save time and cost, we only used the Chinese-English bilingual corpus for the assessment.

4.2 Metrics

To assess translation quality at all levels, we employed the most commonly used BLEU score (Papineni et al., 2002). Additionally, we utilised CHrF (Popović, 2015), TER (Snover et al., 2006) and ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004), and calculated the F1 average of the scores of the ROUGE series to provide a more comprehensive assessment of translation quality at all levels. This is because one indicator may be more sensitive to certain aspects of translation quality, while another may capture different aspects of quality.

4.3 Prompt Construction

Prompts for each level were meticulously crafted to align with the taxonomy's gradable elements. Concretely, regarding the translation styles in Level 2, we added the labels (domain & topic) in the dataset as translation styles to the prompt, such as wikinews(business),

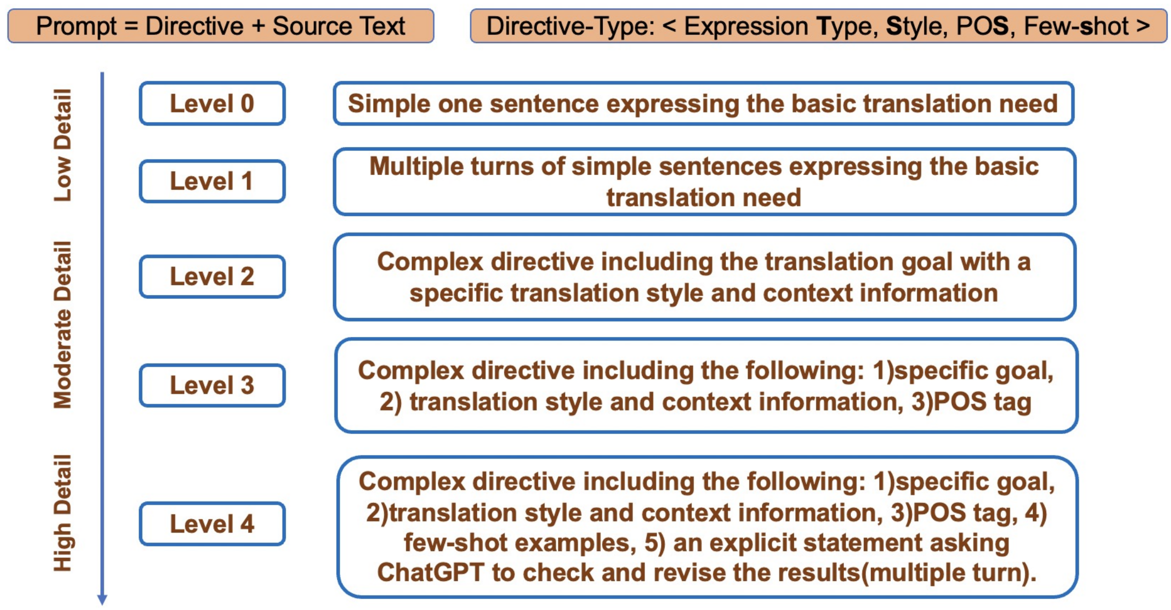


Figure 1: T3S Taxonomy.

Translation	BLEU	CHrF	ROUGE F1(avg)	TER
Level 0	38.42	30.77	0.6132	160.19
Level 1	38.93	31.07	0.6165	146.90
Level 2	40.25	32.41	0.6256	117.12
Level 3	41.25	33.57	0.6303	122.78
Level 4	42.88	36.24	0.6523	112.78

Table 1: ChatGPT's Translation Performance at All Levels.

wikivoyage(travel), wikivoyage(sports), wikibooks(sociology/culture), etc. As for the POS information, we used the open-source natural language processing tool spaCy as our lexical annotation tool to preprocess the source text. With respect to the few-shot examples in Level 4, we randomly selected two sets of source text and target text pairs under the same domain and topic as examples to guide ChatGPT for translation.

4.4 Results

The experimental results, as shown in Table 1, indicate that the translation quality of ChatGPT improves accordingly as the prompt level increases. Specifically, ChatGPT obtains a BLEU score of 38.42 for the basic translation prompt at level 0. This level of prompt is only the most basic translation requirement and does not contain any additional contextual or guidance information. When the prompt upgrades to Level 1, which distinguishes between single-turn and multiple-turn expression types, the BLEU score improves slightly (0.51). However, a more significant increase occurred at Level 2, when the translation style of contextualisation was added to the prompt, which increased the BLEU score to 40.25. This suggests that the inclusion of the translation style has a significant positive impact on the quality of the translation compared to the base translation requirement. At Level 3, the integration of lexical information into the translation prompt further increased the BLEU score to 41.2571. This result suggests that the introduction of lexical information can provide the model with more precise linguistic information, which helps to generate more accurate translations. Finally, the Level 4 prompt achieved the highest BLEU score of 42.88 by including clear instructions, explicit information about the translation context, few-shot examples, and asking the model to check and revise the translation results. Compared to Level 3, the increase was 1.63 points. This largest increase fully demonstrates the remarkable effectiveness of the T3S taxonomy in guiding high-quality translation. Importantly, the Level 4 translation quality is higher than the zero-shot translation quality (BLEU score of 42.50) of GPT-4 under the same dataset and evaluation metrics (Jiao et al., 2023). Moreover, the results for the ChrF, ROUGE series (F1 average), and TER mainly support the above findings.

Overall, the experimental results consistently show that the translation performance of ChatGPT improves significantly as the level of the prompts increases. This suggests that the T3S taxonomy is not only reasonable but also effective in providing targeted guidance for ChatGPT's translation tasks. Future work can explore the application of the T3S taxonomy to more language pairs, as well as the applicability and optimisation space in different domains and scenarios.

5 Case Analysis

This section presents how ChatGPT's performance under the same translation task can be assessed by our T3S taxonomy through a specific translation case. We demonstrate the translation example with different levels of prompts below.

Example Text: MacBook Air is all you — pick your size, pick your color, then go. Whichever model you choose, it's built with the planet in mind, with a durable 100 percent recycled aluminum enclosure. And a fanless design means it stays silent even under intense workloads.

Expert Translation: MacBook Air尽显你的风格，挑个合你的款，选个配你的色，直接搞定。两款机型都采用100％再生铝金属外壳，不管选哪款，都能体现对地球的心意。而无风扇设计意味着即使工作强度高，也能保持静谧无声。

Level 0 Prompt: *Please translate the following text into Chinese:* MacBook Air is all you — pick your size, pick your color, then go. Whichever model you choose, it's built with the planet in mind, with a durable 100 percent recycled aluminum enclosure. And a fanless design means it stays silent even under intense workloads.

Level 0 Translation: MacBook Air 是完全为您打造的— 选择您的尺寸，选择您的颜色，然后开始使用。无论您选择哪个型号，它都是考虑地球环保的，采用了耐用的100％回收铝外壳。而且无风扇设计意味着即使在高强度工作下，它也能保持静音。

Level 1 Prompt (Multiple-Turn form): *First turn: Please translate the following text into Chinese:* MacBook Air is all you — pick your size, pick your color, then go. Whichever model you choose, it's built with the planet in mind, with a durable 100 percent recycled aluminum enclosure. And a fanless design means it stays silent even under intense workloads. *Second turn: Please check and revise the translation results.*

Level 1 Translation: MacBook Air就是为你而设计的— 选择你的尺寸，选择你的颜色，然后开始使用。无论你选择哪个型号，它都是为了地球着想，采用了耐用的100％回收铝外壳。而且，它的无风扇设计意味着即使在高负荷工作下也能保持宁静。

Level 2 Prompt (Advertising style): *Please translate the following text into Chinese in a concise, impressive and advertising style:* MacBook Air is all you — pick your size, pick your color, then go. Whichever model you choose, it's built with the planet in mind, with a durable 100 percent recycled aluminum enclosure. And a fanless design means it stays silent even under intense workloads.

Level 2 Translation: MacBook Air全由您决定— 选择您的尺寸、挑选您的颜色，即刻出发。无论您选择哪个型号，它都以地球为念，采用坚固的100％再生铝外壳打造。无风扇设计确保在高强度工作中也能保持绝对宁静。

Level 3 Prompt (with POS information): *Given the context of an advertisement for an electronic product and the POS tags, please translate this specific sentence into Chinese in a concise, impressive and advertising style:* MacBook Air (Noun) is (Verb) all you (Pronoun) — pick (Verb) your (Pronoun) size (Noun), pick (Verb) your (Pronoun) color (Noun), then (Adverb) go (Verb). Whichever (Determiner) model (Noun) you (Pronoun) choose (Verb), it's (Verb) built (Verb) with (Preposition) the (Determiner) planet (Noun) in (Preposition) mind (Noun), with (Preposition) a (Determiner) durable (Adjective) 100 percent (Percentage) recycled (Verb) aluminum (Noun) enclosure (Noun). And (Conjunction) a (Determiner) fanless (Adjective) design (Noun) means (Verb) it (Pronoun) stays (Verb) silent (Adjective) even (Adverb) under (Preposition) intense (Adjective) workloads (Noun).

Level 3 Translation: MacBook Air就是您的综合选择— 选择您的理想尺寸，选取心仪的颜色，然后出发。无论您选择哪个型号，它都以地球为念，采用坚固的百分之百再生铝质外壳。而无风扇设计确保即使在高强度工作负荷下，也能保持绝对宁静。

Level 4 Prompt (with Few-shot exam-

ples): *First turn: Context Information: It is extracted from an advertisement for an electronic product. Few-shot Examples: 1. Translate "Two perfect sizes. Whether you pick the 13- or 15-inch laptop, both models are superlight and measure just under half an inch thin, so you can take yours anywhere." into "两个尺寸两相宜。13英寸和15英寸两款笔电都十分轻巧纤薄，厚度仅有1厘米多，选哪款都称心称手，去哪里都自由自在。" 2. Translate "Four stellar colors. Your options are out of this world — and each one comes with a matching MagSafe charging cable." into "四款配色都惹眼。每种选择都靓出天际，还配有同色系的MagSafe磁吸充电线。" Considering the context information, few-shot examples and POS tags, please translate the following text into Chinese in a concise, impressive and advertising style:* MacBook Air (Noun) is (Verb) all you (Pronoun) — pick (Verb) your (Pronoun) size (Noun), pick (Verb) your (Pronoun) color (Noun), then (Adverb) go (Verb). Whichever (Determiner) model (Noun) you (Pronoun) choose (Verb), it's (Verb) built (Verb) with (Preposition) the (Determiner) planet (Noun) in (Preposition) mind (Noun), with (Preposition) a (Determiner) durable (Adjective) 100 percent (Percentage) recycled (Verb) aluminum (Noun) enclosure (Noun). And (Conjunction) a (Determiner) fanless (Adjective) design (Noun) means (Verb) it (Pronoun) stays (Verb) silent (Adjective) even (Adverb) under (Preposition) intense (Adjective) workloads (Noun). *Second turn: Please check and proofread the translation to ensure that no errors have been made.*

Level 4 Translation: MacBook Air 轻薄本，是你的不二之选——选尺寸、选颜色，然后出发吧！无论您选择哪种型号，它都专为地球环保而设计，采用100％可回收的铝合金外壳，坚固耐用。而且，无风扇设计确保即使它在高强度工作下也能保持宁静。
As evident from the foregoing, Level 4 provides the closest results to the expert translation. To further validate the efficacy of this taxonomy, we set up both human and machine evaluation methods.

1) We invite three proficient translators to evaluate these five levels of translations according to the quantitative assessment criteria for translation quality proposed by Li (1991), employing the expert translation as the reference standard. This assessment method was set up with four scoring metrics, namely accuracy, fluency, style and textual coherence, with the following weights: W_A=3.5,W_F=2.5,W_S=2,W_C=2. The source text is an advertisement text, which aims at conveying specific messages, emotions and appeals to attract the attention and resonance of the target audience, thus making accuracy and fluency the primary principles of advertisement translation (Xiao, 2010). Advertising's main objective is to promote a product or service, rather than a literature. While style and textual coherence can enhance the appeal and taste of an advert, it should not normally come at the expense of accuracy and fluency. Hence, for this case, the weighting cited as such can provide reasonable constraints on the role played by the scoring indicators in terms of importance and priority.

More specifically, accuracy involves the key purpose of translation, which is to ensure that the translation accurately conveys the message and meaning expressed in the original text. Fluency, on the other hand, emphasises the quality of the written expression of the translation, including regularity, clarity and linguistic fluency. Style stresses the importance of appropriately conveying the stylistic, social and local characteristics of the original text, as well as the extent of the use of rhetorical devices, while maintaining the accuracy of the actual meaning. At last, textual coherence considers whether the arrangement of utterances in the translation adequately takes into account the primary and secondary relationships of the information in the original text, whether contextual co-ordination is achieved, and whether coherence of tone is maintained (Li, 1991).

Each criterion had a maximum attainable score of ten. The ensuing scores represent the respective assessments of the aforementioned trio of professional translators. Moreover, with a view to ensuring the reliability of the results and presenting the evaluation results more explicitly, we calculated the final weighted scores using the average scores of the three translators for all levels of translations under different factors.

As shown in Equation **??** and Table 2, we show the calculation method for the final scores, as well as the different scores and final scores for each level of translation in terms of accuracy, fluency, style, and textual coherence. In the Equation **??**, n stands for the number of professional translators, n=3; W_A, W_F,

W_S, W_C represent the different weights of the four indicators; A_i stands for the specific translator's score for each level of translation in terms of accuracy, F_i stands for the specific translator's score for each level of translation in terms of fluency, S_i refers to the specific translator's score for each level of translation in terms of Style, and Ci denotes the specific translator's score for each level of translation in terms of textual Coherence, i=1, 2, 3.

Based on the score data presented in Table 2, it can be clearly observed that the progression from Level 0 to Level 4 is marked by a discernible trend toward translations that exhibit a greater proximity to expert translation. Specifically, the lowest rated Level 0 (6.8) and Level 1 (7.3) translations score approximately the same in terms of accuracy, and differs significantly in terms of fluency, style and textual coherence. This is due to the fact that their prompts are only different in expression types. Moreover, the improvement (1.0; 0.7; 0.4) in fluency, style and textual coherence from Level 0 to Level 1 proves, to some extent, the effectiveness of multiple-turn prompts in improving the quality of ChatGPT translation results. Compared to Level 2, Level 3 has only a slight improvement (0.2), which is due to the fact that the lexical nature of the source text does not confuse ChatGPT. However, specific texts like legal documents, medical literature, and technical documents often contain specialized terminology, intricate grammatical structures, and polysemous words, and demand a high level of precision and professionalism. In such cases, the inclusion of POS tags becomes crucial as they furnish essential grammatical and semantic information necessary for effectively processing these texts. It should be noted that the addition of POS tags may also increase the preprocessing workload, thus requiring a comprehensive consideration of task requirements and efficiency. Finally, the highest-scoring Level 4 (8.8) translation results exemplify the importance of few-shot examples, which enable ChatGPT to understand the task requirements as well as possible and generate the most brand-specific translations.

2) LLMs not only show excellent capabilities in several NLP tasks such as machine translation, text summarisation, etc., but they are also state-of-the-art translation quality evaluators (Kocmi and Federmann, 2023). Kocmi and Federmann (2023) proposed a GPT-based metric for translation quality assessment, namely GPT Estimation Metric Based Assessment (GEMBA). By conducting experiments on nine versions of GPT models, including ChatGPT and GPT-4, they demonstrated the usefulness and accuracy of pre-trained generative LLMs for translation quality assessment at the system level by using a zero-shot standard prompt. However, this prompt performed poorly at the segment level. Subsequently, Lu et al. (2023b) further validated the capability of LLMs in assessing machine-translated translations. They combined the Chain-of-Thought (CoT) prompting strategy (Wei et al., 2022) and the Error Analysis (EA) paradigm (Lu et al., 2023a) to propose a novel prompting strategy, Error Analysis Prompting (EAPrompt). EAPrompt divides the scoring process into two stages: first, the LLM is prompted to identify the major and minor errors in the translation. The LLM is then asked to count the number of errors in both categories and calculate the final score. Unlike standard prompts, EAPrompt can produce human-like evaluations of machine translations at both the system and segment levels.

Based on such a pioneering discovery, we perform evaluations of all levels of translations in this case with EAPrompt. Below we show the specific prompt template and the scores for each level of translation. Due to space constraints, the Q&A sessions for each level of translation are not presented.

First-turn:
(Source Text)
(Reference)
(Translation)
Based on the given source and reference, identify the major and minor errors in this translation. Note that Major errors refer to actual translation or grammatical errors, and Minor errors refer to smaller imperfections, and purely subjective opinions about the translation.

Second-turn:
Count the number of major and minor errors identified in your last response and compute the final score for this translation. Deduct 5 points for each major error. Deduct 1 point for each minor error. If the translation has no errors, its score will be 0.

	Accuracy			Fluency			Style			Coherence			Final Score
	T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3	
Level 0	7	6	7	7	7	7	6	7	6	7	8	7	6.8
Level 1	7	7	7	8	8	7	7	7	7	8	8	7	7.3
Level 2	9	8	8	8	9	8	8	8	8	8	9	9	8.0
Level 3	8	9	8	9	7	9	8	7	7	9	9	9	8.2
Level 4	9	9	8	9	8	9	9	8	9	9	9	9	8.8

Table 2: Assessment of Accuracy, Fluency, Style and Coherence of Translations at All Levels.

$$FinalScore = \frac{\sum_{i=1}^{n} (W_A * A_i + W_F * F_i + W_S * S_i + W_C * C_i)}{n} \quad (1)$$

	Level 0	Level 1	Level 2	Level 3	Level 4
Results	-27	-23	-22	-18	-12

Table 3: Results of ChatGPT's Quality Assessment of Five Levels of Translations under EAPrompt.

Based on Table 3, it can be concluded that the translation quality exhibits an upward trend as the prompt level increases, which is roughly in line with the results of the human-based translation quality assessment. This result further confirms the effectiveness of T3S Taxonomy and the potential of LLMs in translation quality assessment. However, it is also noted that even higher quality translations (e.g. Level 4 Translation) still received negative scores. This suggests that ChatGPT using EAPrompt may have some rigour in the assessment process or be highly sensitive to subtle differences in translations. This could be due to the fact that the LLMs can capture subtle semantic differences and expressive inconsistencies that may seem acceptable to a human evaluator.

6 Conclusions and Future Directions

This paper highlighted the significance of a taxonomy of prompts for translation tasks, identifying critical design elements such as expression type, style, POS tagging, and few-shot examples. Furthermore, we explored in detail the key roles of gradable translation prompting taxonomy with explicit descriptions and contextual information to enhance the prompts' quality. The synergistic effect of these factors helps to improve translation quality and avoid misunderstanding and ambiguity, which in turn provides more precise guidance for the ChatGPT translation task. Based on the above, we conducted the T3S taxonomy of prompts for ChatGPT translation tasks.

In our study, we evaluated the effectiveness of our translation taxonomy by conducting an experiment using open-source datasets and standard evaluation metrics to rate translation quality across five levels. We also showcased the taxonomy's usefulness through a case study with ChatGPT, highlighting how prompt design impacts translation performance. Our findings offer valuable insights for enhancing ChatGPT's translation applications, prompt optimization, and overall translation quality and efficiency. Building on the insights gathered from our current investigation, future research could focus on comparing ChatGPT, prompted with our refined taxonomy, with dedicated translation services like Google Translate. This would assess whether our taxonomy-based approach can improve the translation capabilities of LLMs to outperform established translation services.

7 Acknowledgements

This work was supported by Gongdong Provincial Key Laboratory of Novel Security Intelligence Technologies (Grant No.2022B1212010005) and XJTLU Research Development Funding (Grant No.RDF-22-01-053).

References

Amplayo, R. K., K. M. Yoo, and S.-W. Lee. 2022. Attribute injection for pretrained language models: A new benchmark and an efficient method. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1051–1064, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chowdhery, A., S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Feng, X., Z. Feng, W. Zhao, B. Qin, and T. Liu. 2020. Enhanced neural machine translation by joint decoding with word and pos-tagging sequences. *Mobile Networks and Applications*, 25(5):1722–1728, oct.

Fu, Y., H. Peng, A. Sabharwal, P. Clark, and T. Khot. 2022. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Gilson, A., C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, D. Chartash, et al. 2023. How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9(1):e45312.

Goyal, N., C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Guo, J., Z. Zhang, L. Xu, H.-R. Wei, B. Chen, and E. Chen. 2020. Incorporating bert into parallel sequence decoding with adapters. *Advances in Neural Information Processing Systems*, 33:10843–10854.

He, Z., X. Wang, Z. Tu, S. Shi, and R. Wang. 2022. Tencent ai lab-shanghai jiao tong university low-resource translation system for the wmt22 translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 260–267, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hendy, A., M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Hlaing, Z. Z., Y. K. Thu, T. Supnithi, and P. Netisopakul. 2022. Improving neural machine translation with pos-tag features for low-resource language pairs. *Heliyon*, 8(8).

Jiang, E., E. Toh, A. Molina, K. Olson, C. Kayacik, A. Donsbach, C. J. Cai, and M. Terry. 2022. Discovering the syntax and strategies of natural language programming with generative language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Jiao, W., W. Wang, J.-t. Huang, X. Wang, and Z. Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Karmaker, S. K. and D. Feng. 2023. Teler: A general taxonomy of llm prompts for benchmarking complex tasks. *arXiv preprint arXiv:2305.11430*.

Kim, G., P. Baldi, and S. McAleer. 2023. Language models can solve computer tasks. *arXiv preprint arXiv:2303.17491*.

Kocmi, T. and C. Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In M. Nurminen, J. Brenner, M. Koponen, S. Latomaa, M. Mikhailov, F. Schierl, T. Ranasinghe, E. Vanmassenhove, S. A. Vidal, N. Aranberri, M. Nunziatini, C. P. Escartín, M. Forcada, M. Popovic, C. Scarton, and H. Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland, June. European Association for Machine Translation.

Kojima, T., S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, X. 1991. Integral optimisation of translations. *Foreign Languages Research*, 04:54–59.

Li, Y., Y. Yin, J. Li, and Y. Zhang. 2022. Prompt-driven neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2579–2590, Dublin, Ireland, May. Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Liu, P., W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Liu, Z., Y. Xu, G. I. Winata, and P. Fung. 2019. Incorporating word and subword units in unsupervised machine translation using language model rescoring. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 275–282, Florence, Italy, August. Association for Computational Linguistics.

Lu, Q., L. Ding, L. Xie, K. Zhang, D. F. Wong, and D. Tao. 2023a. Toward human-like evaluation for natural language generation with error analysis. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5892–5907, Toronto, Canada, July. Association for Computational Linguistics.

Lu, Q., B. Qiu, L. Ding, L. Xie, and D. Tao. 2023b. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint arXiv:2303.13809*.

Pan, W., Q. Chen, X. Xu, W. Che, and L. Qin. 2023. A preliminary evaluation of chatgpt for zero-shot dialogue understanding. *arXiv preprint arXiv:2304.04256*.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Popescu-Belis, A. 2019. Context in neural machine translation: A review of models and evaluations. *arXiv preprint arXiv:1901.09115*.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, and P. Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Press, O., M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12. Association for Machine Translation in the Americas.

Tan, Z., X. Zhang, S. Wang, and Y. Liu. 2022. MSP: Multi-stage prompting for making pre-trained language models better translators. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6131–6142, Dublin, Ireland, May. Association for Computational Linguistics.

Thirunavukarasu, A. J., D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Trautmann, D., A. Petrova, and F. Schilder. 2022. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*.

Wang, B., S. Min, X. Deng, J. Shen, Y. Wu, L. Zettlemoyer, and H. Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada, July. Association for Computational Linguistics.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Xiao, X. 2010. Towards the three principles for translation of english advertising texts. *Journal of Jiangxi University of Finance and Economics*, pages 81–85.

Yang, S., Y. Wang, and X. Chu. 2020. A survey of deep learning techniques for neural machine translation. *arXiv preprint arXiv:2002.07526*.

Yao, S., J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Zhao, Z., E. Wallace, S. Feng, D. Klein, and S. Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Zhou, Y., A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 87-98 recibido 14-12-2023 revisado 15-01-2024 aceptado 29-01-2024

Generation of social network user profiles and their relationship with suicidal behaviour

Generación de perfiles de usuarios de redes sociales y su relación con el comportamiento suicida

Jorge Fernandez-Hernandez,[1] Lourdes Araujo,[1,2] Juan Martinez-Romo[1,2]
[1]NLP & IR Group - Universidad Nacional de Educación a Distancia (UNED)
[2]Instituto Mixto UNED-ISCIII (IMIENS)
{jfernandez, lurdes, juaner}@lsi.uned.es

Abstract: Suicide is one of the leading causes of death worldwide, so characterising individuals with such tendencies can help prevent suicide attempts. In this study, a corpus, called *SuicidAttempt*, of Telegram messaging app users, both with and without explicit mentions of suicide attempts, has been compiled in Spanish. For each user, different demographic features were semi-automatically annotated by different systems, some supervised and some unsupervised. Finally, the collected features and linguistic features extracted from users' messages were analysed to characterise different groups based on their relationship with suicidal behaviour. The results indicate that by detecting these demographic and psycholinguistic features, it is possible to characterise specific at-risk groups and gain detailed insight into the profiles of those who engage in such acts.
Keywords: Suicidal behaviour, profiling, corpus creation, social networks

Resumen: Actualmente el suicidio es una de las principales causas de muerte en el mundo, por lo que poder caracterizar a personas con esta tendencia puede ayudar a prevenir posibles intentos de suicidio. En este trabajo se ha recopilado un corpus, llamado *SuicidAttempt* en español compuesto por usuarios con o sin menciones explícitas de intentos de suicidio, usando la aplicación de mensajería Telegram. Para cada uno de los usuarios se han anotado distintos rasgos demográficos de manera semi-automática mediante el empleo de distintos sistemas, en unos casos supervisados y en otros no supervisados. Por último se han analizado estos rasgos recogidos, junto con otros lingüísticos extraídos de los mensajes de los usuarios, para intentar caracterizar distintos grupos en base a su relación con el comportamiento suicida. Los resultados sugieren que la detección de estos rasgos demográficos y psicolingüísticos permiten caracterizar determinados grupos de riesgo y conocer en profundidad los perfiles que realizan dichos actos.
Palabras clave: Comportamiento suicida, identificación de perfiles, creación de corpus, redes sociales

1 *Introduction*

Suicide is currently one of the leading causes of death worldwide, with approximately 700,000 deaths annually, and the fourth death cause among the young people, as reported by the World Health Organization[1]. In Spain, the number of suicide-related deaths is estimated to be thrice that of those resulting from traffic accidents[2]. Therefore, it is crucial to comprehend the characteristic patterns of individuals with suicidal tendencies and classify them to identify particular social groups with an increased susceptibility

[1]https://www.who.int/news/item/17-06-2021-one-in-100-deaths-is-by-suicide

[2]https://www.mjusticia.gob.es/es/ElMinisterio/GabineteComunicacion/Paginas/211221-NP-Estudio-Epidemiologia-y-Toxicologia-de-las-muertes-por-suicidio.aspx

ISSN 1135-5948 DOI 10.26342/2024-72-6

to suicide.

Characterising population groups with different sensitivities to mental health issues is a crucial step in identifying the most vulnerable populations and designing appropriate support initiatives. Social media platforms offer a wealth of information for this research, as they allow the identification of linguistic and demographic patterns within user-generated content.

This work contributes to the development of a mental health observatory that furnishes health professionals with the most recent data on various population groups, thereby enabling them to accurately interpret individual cases and propose general intervention strategies. The principal objective of this study is to enhance the identification of population profiles associated with suicidal behaviour, using information gathered from social media platforms.

Specifically, a collection of messages regarding suicide attempts on the instant messaging app Telegram has been compiled. The collection has been classified manually, separating positive and negative cases, and has also been annotated with demographic features, such as age, gender, origin or employment status. Semi-automated detection systems, outlined in this article, were utilised for the annotation of these features. These systems are designed specifically for each feature, depending on the availability of external training data and the difficulty of detecting each feature. These systems not only represent a support to manual annotation, but they also serve as the foundation for an automatic system to collect messages and profile the population at risk.

Based on the annotated collection, a study was conducted on the correlation between demographic and linguistic features, and suicidal tendencies. Despite the limited size of the corpus, the findings suggest two key facts. Firstly, there are specific demographic groups that exhibit a considerably higher incidence of suicidal behaviour. In addition, the usefulness of the use of certain linguistic features in at-risk groups has been highlighted.

The article is structured as follows: Section 2 presents the state of the art on the use of artificial intelligence for suicide-related topics, as well as in the profiling of authors. Section 3 presents the collected corpus and the different features annotated in it. Section 4 presents the different methods that have been developed with the aim of extracting the demographic features, and which have been used as an aid to the annotation of the corpus. Section 5 analyses the different features considered in this work that constitute a profile of each user and their relationship with suicide cases. Finally, the conclusions and future work will be presented in Section 6.

2 State of the art

The growing popularity of artificial intelligence has led to its use in more and more fields, including psychology and the identification of different mental disorders. One of the most comprehensive and well-known studies was conducted by Schwartz et al. (2013), which analysed 700 million words and 75,000 volunteers to associate certain words, phrases or speech patterns with different personality profiles. Another recent example is the *CLPsych2019* task (Zirikly et al., 2019), which aims to classify suicide risk into four levels based on Reddit posts written in English. In Du (2023), linguistic features are utilized with classical machine learning methods to predict the most representative psychological state of a text (anxiety, depression, suicide ideation, or "normal"). In (Fernandes et al., 2018), a rule-based system is employed to detect instances of suicidal ideation in English texts, alongside a hybrid approach that utilizes both rules and machine learning techniques to identify suicide attempts. It is also worth mentioning the competition eRisk, which has covered the early detection on the Internet of a wide variety of mental disorders since their first edition in 2017 (Losada, Crestani, and Parapar, 2017). For example, in their last edition the disorders were depression, gambling and eating disorders (Parapar et al., 2023).

The association between distinct demographic features and suicidal behaviour has been a topic of research. Rancāns et al. (2016) conducted a study of the Latvian population and found that middle-aged men living alone and with a low level of education were more likely to exhibit suicidal tendencies, while women with only a low level of education exhibited the highest risk factor. In Akkaya-Kalayci et al. (2018), the study focuses on features associated with personal relationships among young people in Turkey.

The findings suggest that for women, intra-family issues tend to be linked to suicidal behaviour, while for men, relationship problems tend to have a stronger association.

Among the features to be extracted in this work, gender seems to be the one that has been studied the most. Its identification has largely employed classical machine learning techniques, like support vector machines (SVM) or decision trees, rather than deep learning. Among these conventional algorithms, the most successful have been the SVMs (Pizarro, 2019) (Yang et al., 2021), although ensembles have also yielded promising results (Piot-Perez-Abadin, Martin-Rodilla, and Parapar, 2021). Regarding deep learning algorithms, Heidari, Jones, and Uzuner (2020) train separate neural networks for each gender using the Bi-LSTM architecture. Unsupervised learning techniques, such as clustering, can be useful not only for gender identification, but also for analysing the different groups obtained (Bamman, Eisenstein, and Schnoebelen, 2014). On social networks like Twitter, each user has an associated profile picture, which can be used to create classifiers on two levels: on one hand, the images are analysed, while on the other, it focuses on text, with their output being combined. (Wang et al., 2019).

The determination of nationality or provenance has had limited research, with a greater emphasis on handwritten texts rather than text (Al Maadeed and Hassaine, 2014) (Choudhury et al., 2022). Consequently, it is more akin to image analysis than text analysis.

Employment status and profession have typically been addressed as a problem of entity recognition and POS-tagging, as in the case of the MEDDOPROF task (Lima-López et al., 2021). The use of transformers such as *XLM-R* (Lange, Adel, and Strotgen, 2021), the more familiar *BERT* (Mesa-Murgado et al., 2021) or a mixture of the latter with *FLAIR* (Balouchzahi, Sidorov, and Shashirekha, 2021), is the most common method in this scenario.

There is a limited number of studies regarding age, with the majority focusing on *PAN* tasks between 2013 and 2016 (Rangel et al., 2013) (Rangel et al., 2014) (Rangel et al., 2015) (Rosso et al., 2016). These studies predominantly use classical machine learning algorithms such as SVMs or ensembles.

3 Corpus

The corpus was created gathering messages from two different Telegram groups, both focused on mental health problems. One of them was more focussed on suicide, and the other one with focus in anxiety and depression, both groups having positives and negatives users for suicide attempts. These groups are not restricted to a certain nationality, and have users from different Spanish-speaking countries. Despite the fact that the main language is Spanish, some users from non-Spanish-speaking countries can be found, although all of them write in Spanish. So the corpus *SuicidAttempt* comprises 141,894 messages authored by 589 unique users, each user having a mean of 290 messages, between late 2021 and mid-2023 in groups associated with mental disorders.

The users in the corpus can be classified as either positive, where an explicit suicide attempt is mentioned, or negative, where such mention is absent. For classify a user, we first search in his messages for terms related to suicide, with a later manual review to validate. A user who solely mentions suicidal ideation is considered negative. Some examples of sentences that could be considered as explicit mentions are: “I attempted suicide one year ago” or “I consumed thirty pills, but I woke up in the hospital”. Some examples of sentences that could be considered as only ideation are: “I want to die” or “Could anyone give me a quick way to die”.

The 589 users are divided as follows: 156 are positive and 433 are negative users.

Each user in the corpus has received semi-automatic annotation in terms of their gender, origin and employment status. It should be noted that in some cases, users may not provide all information relevant to the traits in question, and therefore no annotation can be made.

The evaluation of the agreement among annotators was measured by Fleiss kappa value (Fleiss, 1971) obtaining 0.78 (“substantial agreement”) in the case of the attempt suicidal annotation, and 0.86 (“almost perfect agreement”) in the annotation of the traits. In simple terms, the kappa coefficient corresponds to the ratio of observed concordances over the total of observations, having excluded all random concordances. The kappa coefficient takes values between -1 and +1.

Three different categories have been defined for the trait "Gender": "**Male**", "**Female**" and "**No binary**", appearing unannotated in case the user in question could not be classified in any of the above three categories during the manual review.

In the corpus, the trait "Origin" was split into two different categories - "Place of birth" and "Place of residence" - because for some users they were different. However, for most of the users both locations will be the same, and will only differ if a specific reference is detected during the manual review.

Something similar occurs with the employment status, which in the corpus is split into two different traits: "Employment status" and "Profession". "Employment status" could be five different categories: "**Work**" for users with job, "**Unemployed**" for those who are unemployed, "**Student**" for those who an explicit mention was found (i.e "I study computer science") or implicitly (i.e "I just got out of class"), "**Homemaker**" for those who have explicitly mentioned their role of homemaker and "**Voluntary**" if an explicit mention exists. If a user does not fit into any of the above categories, then this trait will not be annotated in the corpus.

The other trait is "Profession", which indicates the specific employment activity or place of work. It is possible to find the combination of "**Work**" as an employment status with no profession, e.g. due to a mention of "I just got off work", but without further details about the profession.

Finally, the "Age" trait, instead of being taken as a number, has been divided into the following age ranges: "**<18**", "**18-24**", "**25-34**", "**35-49**" and "**>50**". These ranges are based on those proposed in Rangel et al. (2014), although in our case there were no users over 65, so the highest range is over 50.

An example of the corpus data can be seen in the Table 1.

4 Techniques used for corpus annotation

The annotation process was supported by a series of systems that carried out initial automated tagging, followed by manual revision. The choice of system for each trait relied on the available resources.

In the case of gender, employment status and profession, the systems were supervised; while in the case of place of birth and residence, the approach has been unsupervised. Age was recorded manually, as no dataset with a sufficient number of cases was found to create a system.

4.1 Gender

In the case of gender, there are publicly available datasets that have allowed us to annotate this trait using a machine learning system. A system using the *transformers* technology was also tested but found to be less effective than systems based on classical classification algorithms and thus discarded.

For the base gender detection method, data from the PAN's 2018 and 2019 author profiling task was used (Rangel et al., 2018; Rangel and Rosso, 2019). In both tasks, a dataset of 100 tweets is provided for each user along with their gender information. Users identified as bots were discarded for the 2019 data set. Additionally, messages that were just retweets were also eliminated. The base models were trained on a total of 4,479 users, which were divided as 2,238 female users and 2,241 male users.

For each user, their messages were concatenated using <*FIN*> as tag. Certain special text sequences may appear in the tweets, which have been edited and replaced by different tags:

- The links by <*URL*>
- User mentions (@*username*) by <*USR*>
- Hashtags by <*HTG*>
- The emojis have been removed, although the number of them used by each user has been counted beforehand, to be employed as a feature by the classification algorithms.

The features employed by the classification algorithm can be divided in three different groups:

- **LIWC**: Features obtained from *LIWC 2015* (*Linguistic Inquiry and Word Count*)[3] employing *Spanish Dictionary 2007*. This software identifies 90 dimensions, each one determining the degree that the users employ words that connote positive or negative emotions, self-references, pronouns, etc.

[3] https://www.liwc.app/

Gender	P. Birth	Residence	E. Status	Profession	Age	Suicide
Male	Spain	Spain	Work	Lawyer	35-49	Positive
Female	Spain	Spain	Work	Health	25-34	Positive
Female	Argentina	Argentina	Student		<18	Negative
Male	Colombia	Colombia	Work	English Teacher	25-34	Negative

Table 1: Example of annotation from 4 users in the corpus.

- **TF-IDF**: Features obtained with *Tf-Idf* technique (Term Frequency – Inverse Document Frequency). This method was employed to analyse words, using unigrams and bigrams as terms, as well as characters, using trigrams, tetragrams and pentagrams as terms. In both cases, terms that appeared in over 70% of the documents were excluded. Finally, the Singular Value Decomposition (SVD) technique was applied to reduce the number of features.
- **Number of emojis**: This feature defines the total number of emojis used by a user in their messages.

To obtain the system for semi-automatic annotation of the corpus, firstly the most effective feature mentioned earlier have been found. The SVM method has been employed, as it provided the best results in a preliminary study. The evaluation was carried using a cross-validation with 10 folds. Table 2 shows how the best results are obtained by combining the 3 groups of features (*Tf-Idf* Measure (TFIDF), linguistic features of the *LIWC* (LIWC) and the number of emojis (N_EMO)).

	P	R	F1
LIWC	71.79	71.73	71.71*
N_EMO	65.49	62.61	60.71*
TFIDF	80.93	80.84	80.83*
LIWC+TFIDF	80.97	80.91	80.90*
LIWC+TFIDF+N_EMO	**82.02**	**81.96**	**81.95**

Table 2: Precision (P), Recall (R) and F1-Score (F1) employing SVM and cross-validation with different combination of features for gender identification. A statistical significance test has been carried between the best combination of features and the other options, being significant the difference in all cases (marked with *).

After selecting the optimal combination of features, the next step was to choose the best model. The algorithms tested were: SVM, Decision Trees, Naive-Bayes, Gradient Boosting, Random Forest and AdaBoost, all of them being implementations of the *Sklearn*[4] library, each one trained with their default hyperparameters. As can be seen in Table 3, the best results are achieved with SVM, so this was the model trained on the PAN18 and PAN19 data using the training set. Furthermore, we aimed to assess the model's performance on the test set of the PAN 2019 task, in addition to *SuicidAttempt* corpus, obtaining a precision of 78.55, a recall of 78.39 and a F1-Score of 78.37.

	P	R	F1
SVM	**82.02**	**81.96**	**81.95**
Decision Tree	67.98	67.96	67.96
Naive Bayes	66.79	66.44	66.25
GradientBoosting	79.19	79.15	79.14
RandomForest	77.48	77.45	77.45
AdaBoost	74.48	74.44	74.43

Table 3: Precision (P), Recall (R) and F1-Score (F1) for multiple classification algorithms employing cross-validation for gender identification with Tf-Idf, LIWC and the number of emojis.

The gender of the Telegram users was annotated employing the previous model,and manually revised, achieving the next results (see Table 4): 75.27 of precision, 72.21 recall and 72.56 of F1-Score. The results in this case are slightly lower than those obtained with PAN data. The lower results in the SuicidAttempt corpus may be due to differences with the PAN texts. In any case, the results are high enough to be useful for assisting with the annotation task and can serve as a baseline for future research on gender identification systems.

4.2 Place of birth and residence

The origin of the user is another of the traits considered in this work. In this case, the semi-automatic annotation has not been done based on a supervised learning system, due

[4] `https://scikit-learn.org`

to the unavailability of a dataset that included all the nationalities considered. The methodology employed made use of a dictionary that included all the Spanish-speaking nations (including the Philippines and Equatorial Guinea), Brazil and Portugal as possible origins. For each country, a list of related expressions has been generated, considering the name of the country, the capital, the nationality, and principal cities. For each user, the occurrence of terms associated with a country is counted, and the place with the highest frequency is noted as the origin.

During the manual review, certain users have been found to mention their birth in one country while living in another. For this reason, two different traits were considered: the place of birth and the place of residence.

To derive metrics and perform an analysis on this initial algorithm, we considered an annotation as a hit if it corresponded to the place of birth or the place of residence. This system achieves the next results (see Table 4): a precision of 88.66; 92.46 as recall and 90.33 of F1-Score. These results are sufficient to provide us with a reasonably accurate annotation of the user's nationality.

4.3 Employment status and profession

For the semi-automatic annotation, we used the data from the task *MEDDOPROF* (Lima-López et al., 2021), specifically subtask 2. In this subtask, the objective is to tag professions and identify if they refer to the patient, a sanitary, a familiar or another category.

The data from *MEDDOPROF* have been used, together with the code and process proposed in Lange, Adel, and Strotgen (2021), to train three *transformers* using *xlm-roberta-large* as architecture. One is trained from scratch, while the other two were fine-tuned on the pre-existing models discussed in the article. Once the annotation is done for each of the 3 models, with an ensemble, the results of the three models are combined with a majority vote strategy. Of the referenced entities, only those that have been identified as "PACIENTE" (patient) have been considered. These annotations serve two purposes. Firstly, they serve as input for the employment status identification system. Secondly, they aid in speeding up the annotation of the profession trait.

The annotation of employment status is obtained through a rule-based system using the annotations from the previous ensemble as input. The multiple tagged parts of text, for each user, are reviewed to identify expressions associated with each of the considered employment status, except for "Work". For example, looking for "no" (negation) and "trabajo" (work) in the same sentence to classify a user as "unemployed". If the user has tagged text, but does not meet any of the rules being considered, then the user is classified as "work". If the user has no employment-related mentions, the employment status is left blank. The results of the system can be viewed in the Table 4.

	P	**R**	**F1**
Gender	75.27	72.21	72.56
Origin	88.66	92.46	90.33
E.Status	65.34	60.61	60.87

Table 4: Results obtained by the different systems developed to annotate the traits.

5 *Analysis between the traits and suicidal behaviour*

The study's ultimate aim is to characterise individuals based on their connection to suicidal behaviour. This is done using the demographic traits and the *LIWC* features. It is also important to remember that the corpus is gathered from groups centred around mental disorders, such as anxiety and depression. Thus, these connections are useful for population groups with similar situations, but they may not necessarily be generalizable.

Of the traits noted, gender, age, employment status and place of residence were considered. The place of birth has not been taken into account, as in most cases it is the same as the place of residence and the actual residence has been considered more important to characterize the user. Profession has not been taken into account because it is a category with open labels. In this case, it remains pending, for future work, to standardise this trait in order to be able to analyse it statistically. Users with unidentified traits have been excluded from consideration for each trait. Similarly, labels that appear in less than 5% of users have not been taken into account.

In the case of gender, due to the restrictions mentioned above, only users labelled as

"Male" and "Female" were considered. The results for this trait are presented in Figure 1. The findings reveal a slightly higher frequency of positives among women compared to men, although the difference is small. However, more data is required to determine whether this difference is maintained or increased before concluding that women are a high-risk group.

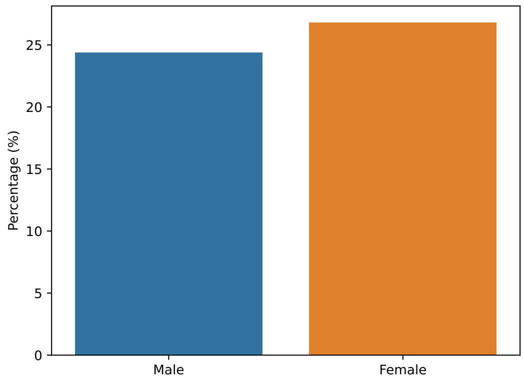


Figure 1: Percentage of positive users for each gender considered.

The next trait studied was age. After applying the above-mentioned restrictions, the ranges *"<18"*, *"18-24"*, *"25-34"*, *"35-49"* were considered. In this case, one can clearly find an at-risk group, the under-18s, with approximately 45% of these users reporting suicidal behaviour, as shown in Figure 2. The following most frequent user group is the 18-24 age range, where almost 40% of users are reported to have attempted suicide. If we divide the age into two groups, under 25 and over 25, and we consider only the positive users, 55% of them were under 25, as shown in Figure 3. Despite the reduced size of the dataset, it seems that age could be a distinguishing factor to consider when examining suicidal behaviour.

Given the high density of users aged below 25 who have attempted suicide, the study examined their employment status to establish any potential correlations. Figure 4 indicates that students have the highest density of positive users, followed by the unemployed.

As it was expected that these groups might be conditioned by age, mainly students, it was decided to study the relationship between both traits. As can be seen in Figure 5, among the users labelled as students, about 90% of them are under 25 years old, which has been observed to be a prevalent group among the positives. Even more interesting is the case of unemployed users, the second group with the highest frequency of positives with values very close to those of students, which in this case are perfectly distributed between those under and over 25 years old. This suggests that the connection between unemployment and suicidal behaviour is more direct and not as age-related as in the case of students.

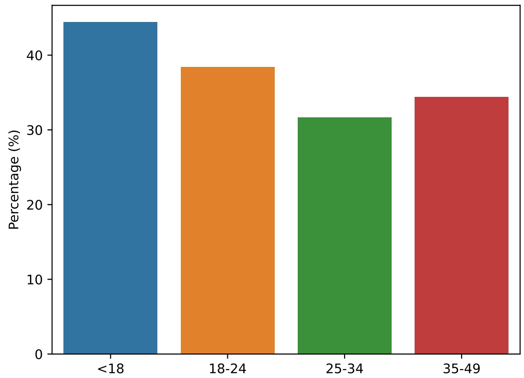


Figure 2: Percentage of positive users in each age group considered.

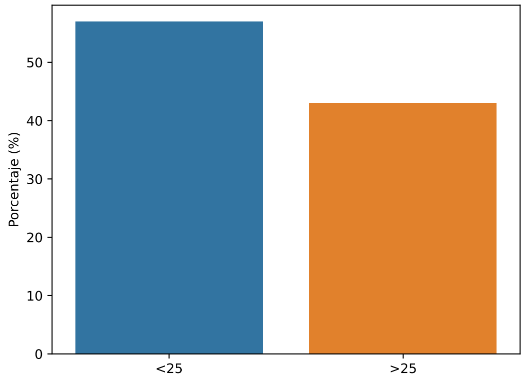


Figure 3: Percentage of individuals under and over the age of 25, from the positive group.

The last demographic trait considered was place of residence, with Spain, Mexico, Argentina, Colombia, Venezuela, and Peru fulfilling the criteria mentioned earlier. Among all of them, it can be clearly seen how Argentina stands out from the rest of the countries, whereas the remaining countries ex-

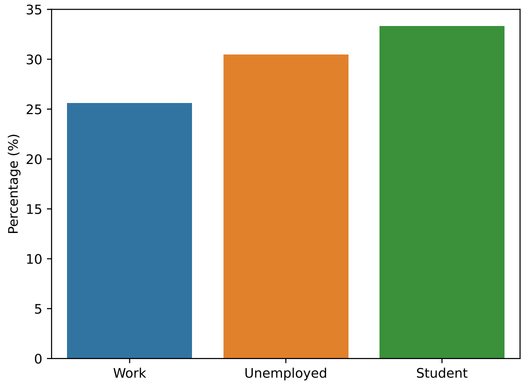


Figure 4: Percentage of positive users for each of the employment status considered.

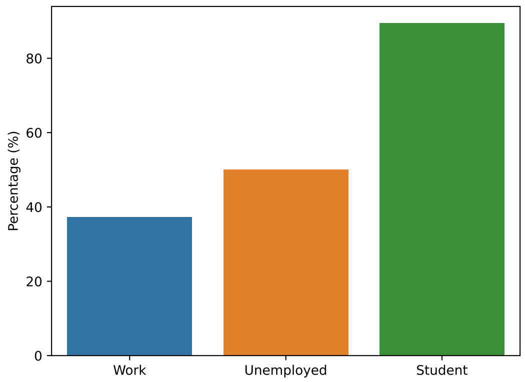


Figure 5: Percentage of individuals under 25, broken down by employment status.

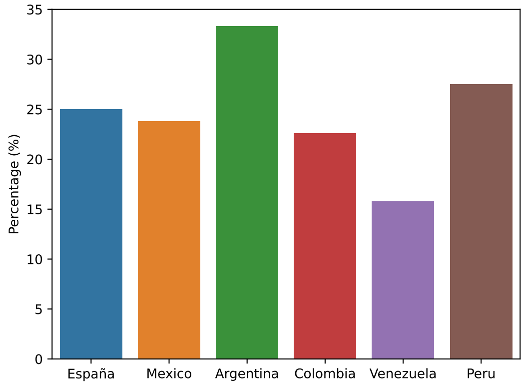


Figure 6: Percentage of suicidal behaviour in the different countries of residence considered.

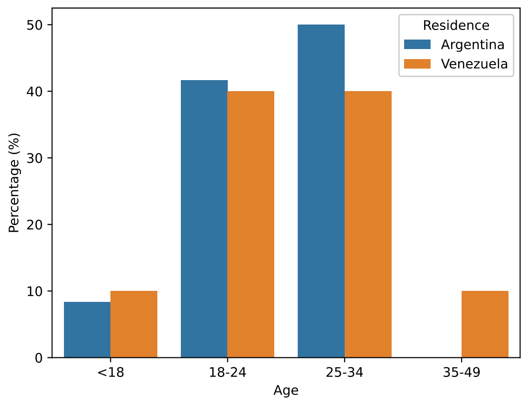


Figure 7: Percentage of age ranges among all users living in Argentina and Venezuela.

hibited comparable proportions, aside from Venezuela, which had a lower frequency, as shown in Figure 6.

To understand the higher suicide attempt rates in Argentina compared to lower rates in Venezuela, we examined age distribution in these countries. In this case, as can be seen in Figure 7, the percentages of age ranges are similar in both nations, with half of all users being under the age of 25. These results seem to indicate that the percentages of positives in these countries would not be conditioned by age.

Other features to consider are the characteristics extracted from the linguistic analyser *LIWC*. These will enable us to characterise users exhibiting positive and negative behaviour based on their linguistic characteristics. By analysing each feature and comparing differences between positive and negative users, we can identify the most significant characteristics (see Figure 8). The feature with the greatest average divergence is *Muerte* (death), reflecting a user's usage frequency of words that have been labelled in the *LIWC* on death-related topics, its usage is higher among positive users compared to negative ones. The subsequent five categories displaying the highest average difference (*Triste* (sad), *Salud* (health), *Enfado* (anger), *Maldec* (cursing) y *verbYO*) mainly relate to negative emotional states or attitudes. The *verbYO* feature, is interesting as it indicates how frequently the first-person singular verb forms are used. Thus, these results seem to indicate that those with suicidal behaviour talk more about themselves. Another interesting feature is the use of negations (feature *Negacio*), which seem to be used on average more by those with suicide attempts.

On the other side, Figure 8, seems to suggest that the negative group uses more words related to cognitive processes according to *LIWC*, in this case the categories *MecCog* and *Certeza* (certainty), as well as an increased use of non-standard punctuation symbols (*OtherP*).

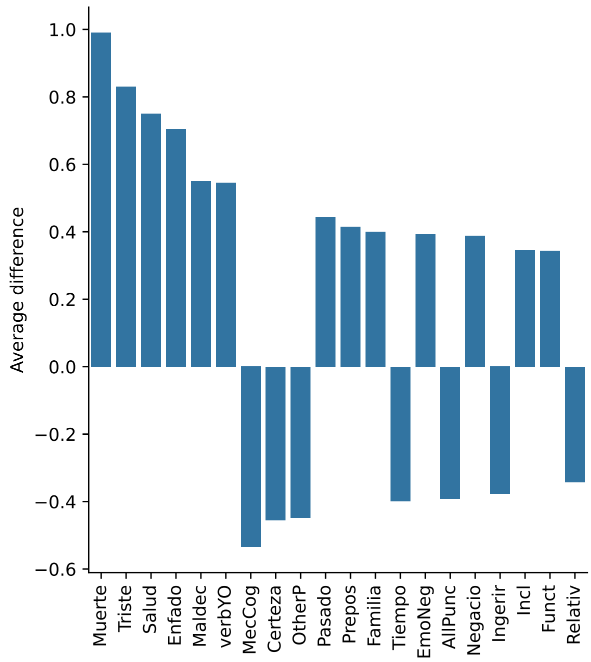


Figure 8: Selection of the 20 LIWC features with the highest average difference. A positive difference suggests higher usage of positive language.

6 Conclusions and future work

In this work, progress has been made in creating a corpus to characterise suicidal behaviour in Spanish. Information on users' gender, place of birth, residence, employment status, profession, and age were recorded. The study also explored different base systems for identifying demographic traits. Lastly, this work attempted to characterise and identify specific risk groups and distinguish linguistic characteristics.

From this analysis it has been observed that the most relevant demographic feature for the study is age, with a higher prevalence among younger users, with around 15% more users under the age of 18 having attempted suicide than in the 25-34 age range. WHO [1] also find prevalence of suicide attempts between the young people.

In terms of gender, a greater proportion of female users have attempted suicide compared to male users, although the variation is not considered significant. Therefore, acquiring further data to determine if this trend remains consistent or if the prevalence of suicidal behaviour among either gender intensifies, would be beneficial.

Our analysis suggests that there is a higher prevalence of suicidal behaviour among students and the unemployed. However, a noticeable correlation with age is evident in the case of the students. Conversely, such a relationship is not observed in the unemployed, a group that is usually associated with economic problems, that has been identified as a risk situation by some organizations, such as the OMS or the WHO. Therefore, it may be beneficial to further explore this demographic group.

Regarding origin, these analyses seem to indicate a higher incidence among people residing in Argentina, but the relationship to age remains unclear. Further investigation of this group is needed to determine if this pattern holds with additional data.

Our conclusions about the linguistic features are similar as the obtained in other studies (Lopez-Castroman et al., 2020). This means that positive users tend to use words that could be classified as negative emotions or feelings, such as sadness or anger, while also frequently using the first person. Negative users, on the other hand, tend to focus on topics that can be framed within different cognitive processes.

Our aim is to extend and advance the current work achieved. Specifically, continue expanding the corpus with more users, and consider other social networks such as Reddit. Furthermore, we will explore additional traits, for example, social or economic issues or addictions.

It would be worthwhile to carry on with the development of automatic extraction systems for various traits, with special attention to age, which could not be determined automatically because there was not a large enough dataset to facilitate this process. The annotations obtained during the development of this work, could be used to develop more sophisticated systems. For example, the counts obtained from each country could be used as features for a machine learning system. Additionally, the best features observed in this work could be combined with the embeddings from a transformer in the case of gender.

Regarding employment status and profession, more work could be done. For example, the annotations of the system, trained with the MEDDOPROF task data, could be used to automatically infer employment status. For the profession, it seems necessary to define categories of professions with similar characteristics, for example, combining doctors and nurses in a category that could be "Health professional". In this way, we will be able to analyse possible relationships between professional groups from different fields and suicidal behaviour.

We also have planned the release of the corpus in the future, under certain commitments. Before the release, we have to study the legal requirements and how to deal with the anonymization of sensitive data such as names.

Finally, it would be interesting to have automatic systems to detect if a user has suicidal behaviour through text, since as seen in Section 5, it is possible to characterise users in this way.

7 Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under Grant PID2019-106942RB-C32, OBSERMENH Project (MCIN/AEI/10.13039 and NextGenerationEU"/PRTR) under Grant TED2021-130398B-C21, and EDHER-MED under grant PID2022-136522OB-C21 as well as project RAICES (IMIENS 2022). The authors would like to thank Hewei Shan for his contribution to the gathering of social media messages for the construction of the corpus.

References

Akkaya-Kalayci, T., N. D. Kapusta, D. Winkler, O. D. Kothgassner, C. Popow, and Z. Özlü-Erkilic. 2018. Triggers for attempted suicide in istanbul youth, with special reference to their sociodemographic background. *International journal of psychiatry in clinical practice*, 22(2):95–100.

Al Maadeed, S. and A. Hassaine. 2014. Automatic prediction of age, gender, and nationality in offline handwriting. *EURASIP Journal on Image and Video Processing*, 2014(1):1–10.

Balouchzahi, F., G. Sidorov, and H. L. Shashirekha. 2021. Adop fert-automatic detection of occupations and profession in medical texts using flair and bert. In *IberLEF@ SEPLN*, pages 747–757.

Bamman, D., J. Eisenstein, and T. Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Choudhury, A. P., P. Shivakumara, U. Pal, and C.-L. Liu. 2022. Eau-net: A new edge-attention based u-net for nationality identification. In *International Conference on Frontiers in Handwriting Recognition*, pages 137–152. Springer.

Du, X. 2023. Lexical features and psychological states: A quantitative linguistic approach. *Journal of Quantitative Linguistics*, pages 1–23.

Fernandes, A. C., R. Dutta, S. Velupillai, J. Sanyal, R. Stewart, and D. Chandran. 2018. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Scientific reports*, 8(1):7426.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Heidari, M., J. H. Jones, and O. Uzuner. 2020. Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 480–487. IEEE.

Lange, L., H. Adel, and J. Strotgen. 2021. Boosting transformers for job expression extraction and classification in a low-resource setting.

Lima-López, S., E. Farré-Maduell, A. Miranda-Escalada, V. Brivá-Iglesias, and M. Krallinger. 2021. Nlp applied to occupational health: Meddoprof shared task at iberlef 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. *Procesamiento del Lenguaje Natural*, 67:243–256.

Lopez-Castroman, J., B. Moulahi, J. Azé, S. Bringay, J. Deninotti, S. Guillaume, and E. Baca-Garcia. 2020. Mining social networks to improve suicide prevention: A scoping review. *Journal of neuroscience research*, 98(4):616–625.

Losada, D., F. Crestani, and J. Parapar. 2017. erisk 2017: Clef lab on early risk prediction on the internet: Experimental foundations. pages 346–360, 08.

Mesa-Murgado, J., P. López-Úbeda, M. C. Díaz-Galiano, M. T. M. Valdivia, and L. A. U. López. 2021. BERT representations to identify professions and employment statuses in health data. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, volume 2943 of *CEUR Workshop Proceedings*, pages 758–769.

Parapar, J., P. Martín-Rodilla, D. E. Losada, and F. Crestani. 2023. erisk 2023: Depression, pathological gambling, and eating disorder challenges. In J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, and A. Caputo, editors, *Advances in Information Retrieval*, pages 585–592, Cham. Springer Nature Switzerland.

Piot-Perez-Abadin, P., P. Martin-Rodilla, and J. Parapar. 2021. Gender classification models and feature impact for social media author profiling. In *International Conference on Evaluation of Novel Approaches to Software Engineering*, pages 265–287. Springer.

Pizarro, J. 2019. Using n-grams to detect bots on twitter. In *CLEF (Working Notes)*.

Rancāns, E., T. Pulmanis, M. Taube, L. Spriņģe, B. Velika, I. Pudule, and D. Grīnberga. 2016. Prevalence and sociodemographic characteristics of self-reported suicidal behaviours in latvia in 2010: a population-based study. *Nordic journal of psychiatry*, 70(3):195–201.

Rangel, F., F. Celli, P. Rosso, M. Potthast, B. Stein, W. Daelemans, et al. 2015. Overview of the 3rd author profiling task at pan 2015. In *CLEF2015 Working Notes. Working Notes of CLEF 2015-Conference and Labs of the Evaluation forum.* Notebook Papers.

Rangel, F. and P. Rosso. 2019. Overview of the 7th author profiling task at pan 2019: bots and gender profiling in twitter. *Working notes papers of the CLEF 2019 evaluation labs*, 2380:1–7.

Rangel, F., P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, and W. Daelemans. 2014. Overview of the 2nd author profiling task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers*, pages 1–30.

Rangel, F., P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. 2013. Overview of the author profiling task at pan 2013. In *CLEF conference on multilingual and multimodal information access evaluation*, pages 352–365. CELCT.

Rangel, F., P. Rosso, M. Montes-y Gómez, M. Potthast, and B. Stein. 2018. Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working notes papers of the CLEF*, 192.

Rosso, P., F. Rangel, M. Potthast, E. Stamatatos, M. Tschuggnall, and B. Stein. 2016. Overview of pan'16: new challenges for authorship analysis: cross-genre profiling, clustering, diarization, and obfuscation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 332–350. Springer.

Schwartz, H. A., J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Wang, Z., S. Hale, D. I. Adelani, P. Grabowicz, T. Hartman, F. Flöck, and D. Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *The world wide web conference*, pages 2056–2067.

Yang, Y.-C., M. A. Al-Garadi, J. S. Love, J. Perrone, and A. Sarker. 2021. Automatic gender detection in twitter profiles for health-related cohort studies. *JAMIA open*, 4(2):ooab042.

Zirikly, A., P. Resnik, O. Uzuner, and K. Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.

Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 99-108 recibido 15-12-2023 revisado 27-01-2024 aceptado 03-02-2024

Enhancing Clarity: An Evaluation of the Simple.Text Tool for Numerical Expression Simplification

Mejorando la claridad: Una evaluación del sistema Simple.Text para la simplificación de expresiones numéricas

Isabel Espinosa-Zaragoza,[1] Paloma Moreda[2] and Manuel Palomar[2]
[1]Centre of Digital Intelligence, University of Alicante
[2]Department of Computing and Information Systems, University of Alicante
isabel.espinosa@ua.es, {paloma,mpalomar}@dlsi.ua.es

Abstract: Numerical information in written texts impacts their readability and is considered complex for people with cognitive disabilities by the Easy-to-Read guidelines. This paper presents Simple.Text, a rule-based system designed to automatically simplify all numerical expressions deemed complex, with a focus on Rules 19-25 from Section 6.2 of the Easy-to-Read guidelines. The results from the evaluation indicate a high precision and accuracy in numerical phenomena detection and transformation, although with some limitations. This system proves to be an efficient and cost-effective tool for the simplification of numerical expressions.
Keywords: Numerical expressions, Easy-to-Read (E2R), rule-based system, cognitive disabilities.

Resumen: La información numérica en los textos escritos afecta a su legibilidad y las pautas de Lectura Fácil las consideran complejas para las personas con discapacidad cognitiva. Este artículo presenta el sistema Simple.Text, un sistema basado en reglas diseñado para simplificar automáticamente todas las expresiones numéricas consideradas complejas, que aborda específicamente las reglas 19-25 de la Sección 6.2 de Lectura Fácil. Los resultados de la evaluación indican una alta precisión y exactitud en la detección de fenómenos numéricos y en su transformación, aunque con algunas limitaciones. Este sistema demuestra ser una herramienta eficiente y rentable para la simplificación de expresiones numéricas.
Palabras clave: Expresiones numéricas, Lectura Fácil, sistema de reglas, discapacidad cognitiva.

1 Introduction

Numerical information in texts impacts their readability (Rello et al., 2013). The simplification of this information is required to guarantee an egalitarian access to information. Facilitating the understanding of language helps citizens to properly exercise their rights and obligations.

The Easy-to-Read guidelines (AENOR, 2018) include several rules for the simplification of numerical expressions, namely Section 6.2, Rules 19-25. These basically entail the clarification of percentages and fractions, dates and hours, and ordinal numbers, amongst others.

As a way of example, percentages like "20%" are transformed into descriptive explanatory clauses in order to avoid using the symbol "%" and providing a more comprehensible quantity. The target audience of these recommendations is people with cognitive disabilities, especially dyslexia and dyscalculia, which are particularly affected by numerical expressions, but not limited to those people.

Previous works emphasise how difficult some numerical expressions are to process and suggest that numbers are more readable in figures than in letters for people with dyslexia (Rello et al., 2013). Additionally, numerical expressions also pose comprehen-

ISSN 1135-5948 DOI 10.26342/2024-72-7

sion problems for people with limited education (Bautista et al., 2011). As can be seen, this issue affects a wide sector of the population that could benefit from more accessible texts. Therefore, Automatic Text Simplification (ATS), “a technology for producing adaptive texts by reducing their syntactic and lexical complexity to make them readable for a user group of users” (Bott and Saggion, 2012), can be of assistance in the simplification of numerical phenomena for any target audience.

The purpose of this paper is to present a rule-based system, Simple.Text, to simplify all of the numerical expressions considered complex in Section 6.2, Rules 19-25, from the Easy-to-Read guidelines. This tool is developed within the ClearText project[1], funded by the MCIN/AEI/10.13039/501100011033 Government and the European Union NextGenerationEU/PRTR (grant reference TED2021- 130707B-I00) and developed by the GPLSI research group[2] of the University of Alicante.

This paper is structured as follows: Section 2 includes a literature review covering ATS tools for numerical expressions; Section 3 presents the papers' objectives and methodology; Section 4 delves into the rule implementation in the system, by describing every category identified and transformed by the system; Section 5 presents the Simple.Text system; Section 6 describes the system evaluation while Section 7 details its findings. Lastly, Section 8 concludes with the future work ahead.

2 Related Work

Previous works in the ATS of numerical expressions are scarce. We depart from the findings from an empirical study in Bautista et al. (2012) on a parallel corpus of original and manually simplified Spanish texts, along with a survey. This study focuses on the simplification of numerical expressions with the intention of implementing the rules computationally, but no actual simplification system is presented.

Similarly, in Drndarević and Saggion (2012), we also encounter the findings of an analysis of a parallel corpus in Spanish (original and simplified) where numerical expressions are taken into account for the development of a simplification system for Spanish. More particularly, (1) the replacing of a word with a figure (“cinco” turns into “5”); (2) the rounding of big numbers (“más de 540.000 personas” turns into “medio millón de personas); (3) the rounding by elimination of decimal points (“1,9 millones” turns into “2 millones”); (4) the simplification of noun phrases containing two numerals in plural and the preposition of by eliminating the first numeral (“cientos de miles de personas” turns into “miles de personas”); (5) the substitution of words denoting a certain number of years (decade, centenary) by the corresponding number; and (6) the representation of thousands and millions in big numbers expressed by means of a word (“17.000” becomes “17 mil”).

To our knowledge, the earliest rule-based system that addresses such issues with a lexical transformation component and a syntactic simplification module is present in Bautista et al. (2013). There we can find a first approximation to the task of simplifying numerical expressions automatically in a text and to varying degrees of difficulty. More specifically, the following replacements are considered: (1) replacing decimal percentages with percentages without decimals; (2) replacing decimal percentages with ratios; (3) replacing percentages with ratios; (4) replacing decimal percentages with fractions; (5) replacing percentages with fractions; (6) replacing ratios with fractions; (7) replacing numerical expressions in words with numerical expressions in digits. This proposal is for English, but with the intention of developing a version for Spanish.

In Bautista and Saggion (2014), the researchers present a rule-based lexical component for the simplification of numerical expressions in Spanish texts based on survey choices for simplification. The system is composed of: (1) text processing using FreeLing; (2) the transformation of the FreeLing output into XML representation; (3) the application of grammars for numerical expression recognition; (4) the simplification of target numerical expression; and lastly, (5) a sentence rewriting stage. Among the numerical expressions tackled in this work, there is the rounding of percentages (“18,55%” is transformed into “19%”) but not the simplification of the percentage in itself, as recommended by the European Easy-to-Read

[1] `https://cleartext.gplsi.es/`

[2] `https://gplsi.dlsi.ua.es/`

guidelines (AENOR, 2018).

Lastly, an ATS system for Spanish is presented in Bautista et al. (2017), where the following phenomena are considered: (1) partitive numerals like for example, "un millón" (a million) or "una centena" ("a hundred"); (2) monetary expressions consisting of quantity and the monetary unit, as in " 2.000 dólares" ("2,000 dollars"); (3) fractions and percentages, like "34%", are substituted with the lemma "34/100"; and (4) physical measures, for example, "30 km/h".

As can be observed, apart from the scarcity of systems for the simplification of numerical expressions in Spanish, the papers presented offer a partial and not a global solution to the simplification of numerical expressions. That is, these do not encompass the entire range of numbers identified as obstacles in the Easy-to-Read guidelines. Thus, we propose a rule-based system to tackle the entirety of these numerical phenomena described in the Easy-to-Read guidelines (AENOR, 2018).

3 Objective and Methodology

The objective of this paper is to identify and resolve the complexity associated with the numeric phenomena deemed as difficult to comprehend by the Easy-to-Read guidelines (AENOR, 2018) in a rule-based system by transforming them into simpler expressions.

As the Easy-to-Read guidelines are often general and flexible rules, the collaboration with the non-governmental organisation APSA[3] has enlightened the path by defining the restrictions to such rules. This NGO has a group of expert Easy-to-Read validators with cognitive disabilities. From this collaboration, we have been able to define and restrict several rules that were rather loose. For example, Rule 19 in Section 6.2 suggests using Arabic numerals. However, for the numbers "100" and "1000", APSA recommends using the written version (e.g. "cien" (a hundred) and "mil" (a thousand), respectively). Drawing on the expertise of APSA's specialists in text simplification and validation according to the Easy-to-Read guidelines, we opted to incorporate their insights and rule specifications into our system. This decision was made to leverage their expertise on the matter, contributing to the creation of a more effective system thanks to this synergy.

Our methodology consists of the following steps:

- 1. Developing a system for the identification and transformation of numerical expressions in Spanish texts.
- 2. Building a rule-based system for automatic simplification of numerical expressions.
- 3. Evaluating the automatically simplified output.

4 Rule Implementation

All the rules contemplated in this system correspond to Section 6.2, the lexical simplification section, in the Easy-to-Read guidelines (AENOR, 2018). Many rules include more than one transformation or implementation. Table 1 includes a summary of each rule implemented by the system and its corresponding rule number plus an example.

First of all, it is necessary to identify numbers in the text whether they are expressed in letters or figures. For this reason, it is imperative to undertake a preliminary processing in order to identify and resolve numbers in letters. In this preprocess, numbers in letters are identified using SpaCy library.[4] Once a number written in letters is identified, then it is replaced by its corresponding Arabic numeral using a predefined dictionary. Then, the process to identify and resolve numbers is run.

The process is divided into two phases: (i) identification and (ii) resolution. The identification phase is carried out using the SpaCy tool (part-of-speech tagging) in order to determine if a word begins and finishes with numerical characters. In that case, we will regard it as a number to be treated in the resolution phase.

The resolution phase requires different converting rules depending on the type of number identified. Consequently, it is essential to determine to which numerical category each of the numbers identified in the identification phase belongs to. This is accomplished in the following order: dates, times, telephones, percentages, ordinals and Roman numerals and other quantities. In this way, in order to assign a number to a category, we first verify that it has not been identified

[3] https://www.asociacionapsa.com/

[4] https://spacy.io/

Rule	Numerical expression	Original	Easy-to-Read
Rule 19	Figures	dos	2
Rule 19	Rounding quantities	139	más de 100
Rule 19	Explain big numbers	60.000	60 mil
Rule 20	Phone numbers	123456789	123 45 67 89
Rule 21	Ordinal numbers	primero, undécimo	primero, 11
Rule 22	Percentages	20%	2 de cada 10
Rule 23	Dates	01/01/(20)20, 01-01-(20)20	1 de enero de 2020
Rule 24	Time	23:30	11 y media de la noche
Rule 25	Roman numbers	Jaime I	Jaime primero

Table 1: Summary of the rules implemented.

in any of the previous categories, and then we check if it complies with the identification rule of that particular category. This sequential order is necessary to prevent incorrect and duplicate substitutions.

The following subsections include the different numerical expressions or categories considered complex for people with cognitive disabilities by the Easy-to-Read guidelines. Each category is defined, followed by an explanation of its detection pattern and resolution or transformation process.

4.1 Dates

According to the Easy-to-Read guidelines, compact dates expressed with hyphens or slashes are not recommended. These include orthotypographic symbols that can be complex to process (Section 6.2, Rule 23, (AENOR, 2018)). This recommendation closely aligns with Rule 8 in Section 6.1 (AENOR, 2018), which specifies that these orthotypographic symbols should be avoided.

The identification of dates is achieved by using a regular expression. By utilising this regular expression, dates following the format DD/MM/YY(YY) are identified. It must be pointed out that the year can be two or four digits and expressed by means of a hyphen or a dot as a separator.

To transform the detected dates into the recommended format, a dictionary is utilised to establish the relationship between months in letters and months in numbers. Hence, dates such as “12/04/2020” or “12-04-20” should be displayed fully written, as follows: “12 de abril de 2020”.

4.2 Times

To identify whether a number represents a time, a regular expression is used to detect the format “hh:mm”. When it comes to the automatic simplification of time, time slots are highly cultural. In Spanish these are:

- In the morning (from 06:00 to 12:59), e.g. 6 y media de la mañana.
- In the afternoon (from 1:00 p.m. to 8:59 p.m.), e.g. 1 y 20 de la tarde.
- At night (from 9:00 p.m. to 12:59 a.m.), e.g. 6 y 35 de la tarde.
- In the early morning (from 01:00 to 05:59), e.g. 3 menos cuarto de la madrugada.

Therefore, set hours (e.g. o’clock, quarter past, half past and a quarter to) are always written (e.g. en punto, y cuarto, y media, menos cuarto). Consequently, instead of “23:30 PM”, the simpler written version should be “11 y media de la noche”. In-between times, like “10:10 AM” or “23:35 PM” should be transformed into “10 y 10 de la mañana,” and “11 y 35 de la noche”, respectively.

4.3 Telephone Numbers

If the identified number in the text consists of nine consecutive digits, and the sentence in which the number appears contains the word “teléfono” (telephone) or “móvil” (mobile) either two words before or after the nine consecutive digits, it will be classified as a telephone number. In such instances, the correct transformation resolution is to separate the nine consecutive digits with spaces following a 3-2-2-2 structure (e.g. 123 45 67 89).

4.4 Percentages

Concerning the simplification of percentages, it is recommended to find alternative rephrasing options, as orthotypographic symbols such as “%” are regarded as complex (i.e. Section 6.1., Rule 8, (AENOR, 2018)). Thus, if the character after the number is the percentage symbol (%), then the number is categorised as a percentage (e.g. 20%, 37%).

Once identified, it is then substituted with an analogous expression using a rule that replaces the number and the percentage symbol with the number and a text. This text varies depending on whether the amount is divisible by 10 or not. On the one hand, when the number is divisible by 10, the text is “ x de cada 10” (x out of 10). On the other hand, when the number is not divisible by 10, the text is “x de cada 100” (x out of 100). this is exemplified below:

- “20%” is simplified as “2 de cada 10” (2 out of 10).
- “37%” is simplified as “37 de cada 100” (37 out of 100).

4.5 Ordinal Numbers

In line with the Easy-to-Read guidelines, the use of ordinal numbers should be changed to cardinal. Nevertheless, our collaborators affirm that written ordinal numbers from one to ten are understood by people with cognitive disabilities and do not need to be changed, according to their experience. For example, “primer/primero(s)/a(s)” (first), “segundo(s)/a(s)” (second), “tercer, tercero(s)/a(s)” (third), etc. Thus, these remain ordinal and in written form, as the validators understand them. However, from eleven onwards, these are changed to cardinal numbers: “undécimo” (eleventh) is changed to “11”. Therefore, an example such as “Juan vive en la planta 18º/decimoctava” (Juan lives on the 18th/eighteenth floor) should be reworded to a simpler version using cardinal numbers, like the following: “Juan vive en la planta número 18” (Juan lives on floor number 18). The detection and substitution are performed by using a dictionary that has been manually created specifically for this purpose.

4.6 Roman Numerals

As for the named entities when these are proper names of kings, Roman numerals adapt to letters. Nevertheless, they do not undergo a double adaptation from Roman numerals to letters and from ordinal numbers to cardinals in the first ten cases:

- “Jaime I” (James I) changes to “Jaime primero” (James the first).
- “Siglo XX” (20th century) becomes “Siglo 20” (20 century).

Even though the Easy-to-Read guidelines (AENOR, 2018) indicate including “que se lee” (what reads as) when treating these cases (e.g. Alfonso X que se lee Alfonso décimo), our collaborators indicated that it is much more straightforward to do it in this way. In order to detect and replace Roman numerals a Roman Phyton Library[5] is used.

Within the Roman numerals, it is necessary to distinguish kings’ names (e.g. Jaime I) since the resolution process is different. This case is identified when Named Entity Recognition (NER) + Roman numeral appears in the text. Upon detection using SpaCy, the Roman number is replaced by the corresponding ordinal number (e.g “Jaime I” becomes “Jaime primero”).

4.7 Other Quantities

If the detected number has not been considered in any of the above categories, then it is regarded as a quantity. To identify them, a regular expression is applied to detect numbers with or without (1) a thousands separator or (2) a decimal point for decimals. The substitution process is subdivided into the following subsections:

4.7.1 Figures

It is recommended to write numbers in figures up to 1,000, that is, from 1 to 999. This is already done in the preprocessing phase explained in Section 4.

4.7.2 Explain Big Numbers

From there, big numbers are changed to a hybrid format where part of the number is written with Arabic numbers and the rest is expressed in written format: “2 mil” (2 thousand). This approach replaces the zeros with their textual equivalent, rather than representing them as numerals. This complies with the Easy-to-Read guidelines (AENOR, 2018), which state that numbers with many digits are difficult to read and, thus, writing them in letters can make them easier to understand. To facilitate their understanding, alternative options are contemplated, such as:

[5] https://pypi.org/project/roman/

- Qualitative comparisons (e.g. as many people as those who live in Granada).

- Replacement by terms such as " several", " thousand" and others when the context allows it.

When it comes to numbers, what is considered big is open to interpretation. The Easy-to-Read guidelines are flexible in so much that these do not set a limit in principle: it depends on the validation sessions and whether it is understandable or not there. Most of the time it depends on the context and the relevance that this number in question has in the text. Currently, according to the validation groups working on this project, we established that the number from which we would apply this is "10.000" which is transformed into "10 mil" (10 thousand). That being said, "100" and "1.000" are also transformed into "cien" (a hundred) and "mil" (a thousand), as previously discussed in Section 3.

4.7.3 Rounding Quantities

Rounding numbers is recommended by the Easy-to-Read guidelines (AENOR, 2018) at the expense of losing precision. This is applied to decimal numbers (e.g. "1.3" is rounded to "1") and other quantities. That is, "1.999" is rounded to "casi 2 mil" (almost 2 thousand). Nevertheless, some exceptions are contemplated, like ticket prices, won prizes, and others, although no implementations are applied yet in this regard until we enter the project's meaning and disambiguation module.

5 *The Simple.Text System*

The current version of the Web App allows for the selection of (1) individual language phenomena simplification, enabling the simplification of specific language phenomena such as superlative forms or *-mente* adverbs, amongst others; (2) language level simplification, which offers the choice of simplifying the entire palette of linguistic phenomena organised by language levels (currently limited to lexical and syntactic); and (3) applying all simplifications at once. Subsequently, users submit the text for simplification on the top box and obtain the output in the box below. Figure 1 illustrates an example of simplification in the current preliminary interface.

6 *System Evaluation*

The system evaluation is performed by detecting and resolving the numeric phenomena in 5,000 texts from the CLEARSIM corpus, which contains texts from the public administration. This accounts for one third of the total texts in that corpus. These texts were gathered from the official websites of municipalities in the Alicante area, focusing on the domains of culture, sports, and leisure. We utilised the Simple.Text System to identify and transform the numeric expressions deemed complex by the Easy-to-Read guidelines (AENOR, 2018).

Given the impossibility of presuming complete system detection and transformation, we conducted a manual evaluation involving a representative quantity of texts to simplify and, subsequently, we scaled the results. This corpus will be available on the project's website.[6] To do so, we extracted a representative number of texts out of the 5,000 texts by following the Formula 1 presented in (Pita-Fernández, 1996):

$$M = \frac{N * K^2 * P * Q}{E^2 * (N-1) + K^2 * P * Q} \quad (1)$$

The symbols in the equation stand for the following: N for population, K for the confidence interval, P for the success probability, Q for failure probability and E for the error rate. The values given to each of these parameters, more specifically, K=0.95, E=0.05, P=0.5, and Q=0.5 were taken from (Vázquez et al., 2010).

After calculating the formula, the resulting number of texts M was 89, which then was rounded up to 90 texts. These texts were manually analysed by a human to check the accuracy of both the linguistic phenomena detection and the linguistic phenomena resolution. The human detection evaluation yielded 1,597 numerical expressions that are categorised as follows:

- Figures: 966
- Written numbers: 178
- Decimal numbers: 51
- Dates: 14
- Hours: 226
- Percentages: 11

[6] `https://cleartext.gplsi.es/`

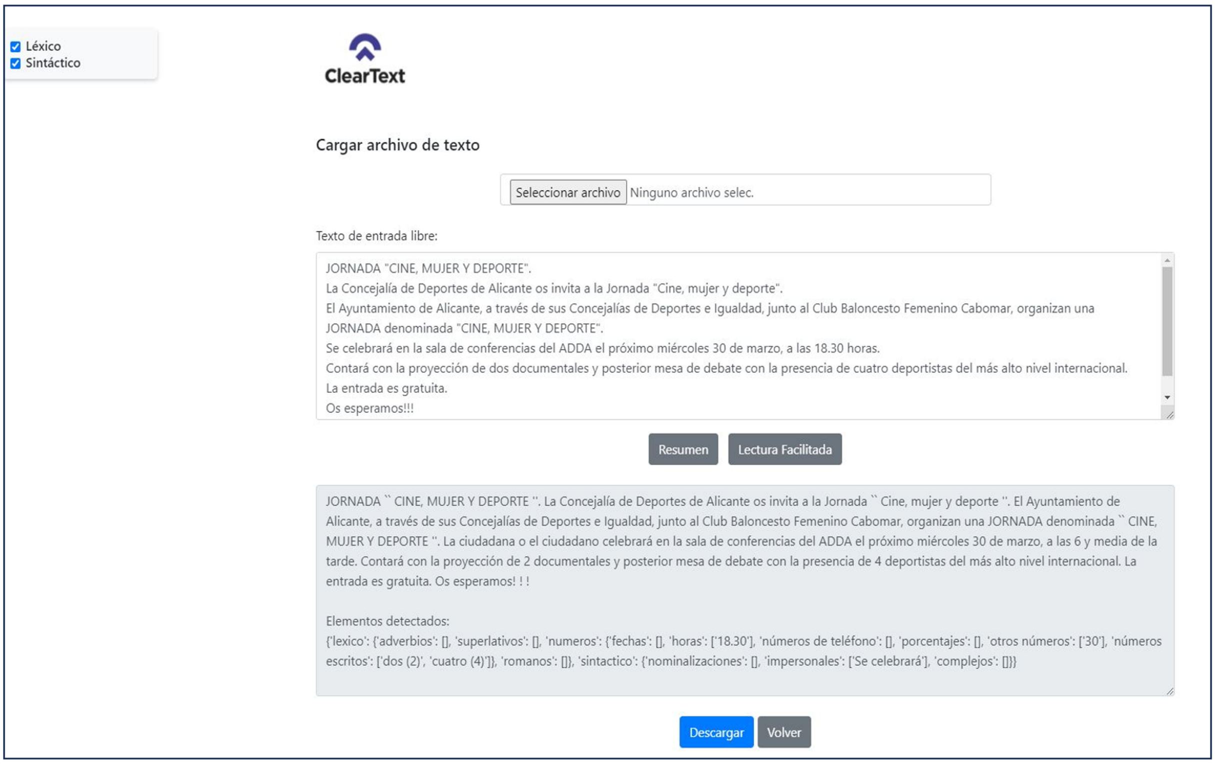


Figure 1: Simple.Text Tool.

- Ordinal numbers (both written and in figure): 92
- Roman numbers: 53
- Phone numbers: 6

After this initial human detection, an evaluation of the system's detection and transformation was performed.

7 *Discussion of Results*

This section discusses both the detection and transformation of numerical expressions in the current version of the Simple.Text system.

The results for the detection of numerical phenomena are presented in Table 2, which includes a description of the accuracy, precision, recall and F1-score (Derczynski, 2016) for the detection of every single numerical expression analysed. Regarding the detection of the numerical categories, we observe an overall good detection except for telephone numbers, dates, times, ordinal numbers and Roman numbers. Some of these issues are caused due to the different ways in which authors express these phenomena in the text. For example, telephone numbers separated with a full stop (e.g. 123.456.789) or in between hyphens (123-45-67-89) were not identified. Similarly, telephone numbers correctly written according to the rules in the original texts were not identified as such, but as quantities. Dates expressed in the format DD.MM.YYYY were not detected and times with the abbreviation h (hours) adjacent to the last numeral character prevented the identification of times (e.g. 07:00h). Text 1043 is a representative example with 22 cases of times expressed in this way but not detected. Similarly, quantities followed by symbols such as €, *km*, etc. prevented the detection of such figures. Roman numbers is the only category with precision below 1. This happened due to the detection as Roman numbers of entities that were not numbers (e.g. the abbreviation *CC*, "Centro Comercial", meaning "shopping centre" was identified as a Roman number). This could be counteracted with the dictionary covering abbreviations, which is a step we will potentially take in the near future.

Concerning the transformation of the categories (see Table 3), all of them perform correctly (e.g. 1) except one quantity that is not rounded (e.g. 1.125) and 16 Roman numbers that are transformed in an incorrect way (e.g. Jaime II as Jaime 2 instead of Jaime segundo). Both of these are system errors and

Category	Accuracy	Precision	Recall	F1
Figures	92.96	1	92.96	96.35
Rounding quantities	98.03	1	98.03	99.01
Explain big numbers	96.62	1	92.62	98.28
Dates	0.5	1	0.5	66.66
Times	60.61	1	60.61	75.48
Percentages	1	1	1	1
Ordinal numbers	66.30	1	66.30	79.73
Roman numbers	70.66	70.66	1	82.81
Telephone numbers	16.66	1	16.66	28.57

Table 2: System evaluation, data detection. All the data expressed in percentages.

Category	Accuracy	Precision	Recall	F1
Figures	99.88	1	99.60	99.88
Rounding quantities	1	1	1	1
Explain big numbers	1	1	1	1
Dates	1	1	1	1
Times	1	1	1	1
Percentages	1	1	1	1
Ordinal numbers	1	1	1	1
Roman numbers	69.81	1	69.81	82.22
Telephone numbers	1	1	1	1

Table 3: System evaluation, data transformation. All the data expressed in percentages.

this evaluation will help us fix these issues in a later system version.

Another issue we encountered is the fact that we need more context or meaning to determine if, for instance, "1999", is a quantity or a year. Out of 198 correct roundings, 114 were years and not quantities. Therefore, 57,57% of correct figure transformations are technically not correct with respect to the text. It remains imperative to establish a method for resolving ambiguity in such instances in future meaning and disambiguation modules in the project.

In that regard, previous works already highlight the importance of simplifying taking into account the local context of the sentence (Bautista and Saggion, 2014). For instance, in a context where a comparison is taking place, if rounding is applied, no information will be transmitted. See the example provided by the authors: "The numbers of dissolutions are maintained at 2010 similar to those of 2009, 22,435 versus 21,875, with a slight increase of 2.56%" (Las cifras de disoluciones se mantienen en 2010 similares a las de 2009, 22.435 frente a 21.875, con un ligero incremento del 2,56%.). This case puts in the forefront the fact that regular expressions, which disregard context, have their shortcomings in cases such as the one exemplified. Thus, syntactic awareness is key to avoiding simplification fails. All in all, although limiting the scope of the transformed phenomena, we ensure that the transformations are correct with a rule-based system.

8 Conclusions and Future Work

The main contribution of this research is the implementation of the entirety of Easy-to-Read guidelines dealing with numbers to a rule-based system (i.e. Simple.Text), within the context of the Clear.Text project. The advantages of rule-based systems lie in their precision and ability transform accordingly with a very cost-effective approach. While we value the use of Large Language Models (LLMs), we understand that these are not strictly necessary for clear-cut and well-defined specific tasks. When compared to an LLM, with this approach we gain explainability and resources.

With the rule knowledge that we have gathered for the simplification of numerical entities, we could define two tasks to solve in the future: to identify both (1) the numerical entities and (2) their category in a given text, which directly refers to the transformation that it should undertake for its resolution and therefore, its simplification. Then, instead of having a set of rules that are limited by its disconnect to context, we could build a corpus to train a machine learning (ML) model that infers these rules. In this way, explainability would not be sacrificed, as many traditional ML models offer explainability.

Overall, we could improve the tool by creating a hybrid system where the detection and classification could be performed with machine learning, deep learning or even BERT, and the transformation phase to be performed with rules, which ensures a precise and accurate transformation. In this way, we would not need a large simplification corpus to train a LLM.

Future work also includes the refinement of the system's current rules, the continuation of the implementation of the entirety of Easy-to-Read guidelines and the evaluation of the system with control and cognitive disabled groups.

More specifically, regarding percentages, there are some exceptions that will be treated using an ad hoc dictionary specifically created for that purpose, for example, "50%" will be replaced by "la mitad" (half), as our collaborators indicate that this construction is easier to comprehend than "5 out of 10". Similarly, fractions will be treated and solved as percentages, that is, with constructions that transmit the same information, for example, "uno de cada tres" (one out of three) instead of "1/3".

Groups of numbers represented in one word, such as "decena" (ten), "docena" (dozen), "millar" (thousand), "centena" (hundred), "centenario/a" (centenarian) or "milenario/a" (millennial), among others, could also be difficult to comprehend. Although these are not explicitly acknowledged in Section 6.2, Rules 19-25 in the Easy-to-Read guidelines (AENOR, 2018), they could be addressed in future work related to numerical expressions using a dictionary.

The resources created by this project will be available on Huggingface[7] and the research group's GitHub[8], as well as the official webpage of the project.

[7] https://huggingface.co/gplsi
[8] https://github.com/gplsi

Acknowledgements

This research was conducted as part of the ClearText project (TED2021- 130707B-I00), funded by the MCIN/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR. Additionally, we acknowledge the collaboration of COOLANG.TRIVIAL: Technological Resources for Intelligent VIral AnaLysis (PID2021-122263OB-C22) funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe" as well as SOCIALFAIRNESS.SOCIALTRUST: Assessing trustworthiness in digital media (PDC2022-133146-C22) funded by MCIN/AEI/10.13039/501100011033/ and by the "European Union NextGenerationEU/PRTR".

References

AENOR. 2018. Norma española experimental une 153101 ex. lectura fácil: Pautas y recomendaciones para la elaboración de documentos.

Bautista, S., B. Drndarevic, R. Hervás, H. Saggion, and P. Gervás. 2012. Análisis de la simplificación de expresiones numéricas en español mediante un estudio empírico. *Linguamática*, 4(2):27–41.

Bautista, S., R. Hervás, P. Gervás, R. Power, and S. Williams. 2011. How to make numerical information accessible: Experimental identification of simplification strategies. In *Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part I 13*, pages 57–64. Springer.

Bautista, S., R. Hervás, P. Gervás, R. Power, and S. Williams. 2013. A system for the simplification of numerical expressions at different levels of understandability. In L. Rello, H. Saggion, and R. Baeza-Yates, editors, *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 39–48, Atlanta, Georgia, June. Association for Computational Linguistics.

Bautista, S., R. Hervás, P. Gervás, and J. Rojo. 2017. An approach to treat numerical information in the text simplification process. *Universal Access in the Information Society*, 16:85–102.

Bautista, S. and H. Saggion. 2014. Can numerical expressions be simpler? implementation and demostration of a numerical simplification system for Spanish. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 956–962, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Bott, S. and H. Saggion. 2012. Automatic simplification of Spanish text for e-accessibility. In *Computers Helping People with Special Needs: 13th International Conference, ICCHP 2012, Linz, Austria, July 11-13, 2012, Proceedings, Part I 13*, pages 527–534. Springer.

Derczynski, L. 2016. Complementarity, f-score, and nlp evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 261–266.

Drndarević, B. and H. Saggion. 2012. Towards automatic lexical simplification in Spanish: an empirical study. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 8–16.

Pita-Fernández, S. 1996. Determinación del tamaño muestral. *Cadernos de atención primaria*, 3(3):138–141.

Rello, L., S. Bautista, R. Baeza-Yates, P. Gervás, R. Hervás, and H. Saggion. 2013. One half or 50%? an eye-tracking study of number representation readability. In *Human-Computer Interaction–INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part IV 14*, pages 229–245. Springer.

Vázquez, Y. G., A. F. Orquín, A. M. Guijarro, and S. V. Pérez. 2010. Integración de recursos semánticos basados en WordNet. *Procesamiento del lenguaje natural*, (45):161–168.

Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 109-121 recibido 15-12-2023 revisado 11-01-2024 aceptado 31-01-2024

Del discurso a la acción: Clasificación de actos de habla en textos legislativos

From Discourse to Action: Classification of Speech Acts in Legislative Texts

Doaa Samy
Cairo University, Egypt
Universidad Complutense de Madrid, España
doaasamy@cu.edu.eg
dkhalil@ucm.es

Resumen: Los actos de habla son unidades básicas de la comunicación lingüística que permiten realizar acciones a través del lenguaje. En el texto legislativo, los diferentes tipos de actos de habla cobran especial relevancia a nivel pragmático porque detrás del lenguaje, existe una intención (acto ilocutivo) que va más allá de las palabras para organizar y cambiar la realidad en una sociedad. La teoría lingüística propone diferentes tipos de actos de habla. Este trabajo tiene como objetivo clasificar automáticamente tres tipos por su relevancia en el texto legislativo: 1) Los actos asertivos que describen hechos y realidades; 2) los actos directivos que definen las normas o regulan las relaciones y las competencias de la materia en cuestión; y 3) los actos compromisorios que reconocen los derechos y se comprometen a velar por estos derechos. Para la clasificación, se ha anotado un conjunto de 1325 enunciados divididos en subconjuntos de entrenamiento, validación (80%-20%) y un conjunto de prueba (250 enunciados). Se han entrenado y se han evaluado varios clasificadores automáticos multi-etiqueta y multiclase basándose en tres tipos de modelos: modelos clásicos de aprendizaje automático, modelos fundacionales del lenguaje (*LLMs*) de tipo "encoder" y un modelo fundacional generativo de tipo "decoder" mediante instrucciones *prompting* de 5 niveles (GPT 3.5). Los clasificadores basados en modelos "encoder" (*BERT* y *RoBERTaLex*) han obtenido los mejores resultados. *BERT* ha alcanzado un *f1-macro* de 0,85 y un *f1- micro* de 0,87. *RoBERTaLex* ha obtenido 0,86 en *f1-macro* y *f1-micro*.
Palabras clave: Modelos fundacionales, actos de habla, pragmática computacional, procesamiento de textos legales.

Abstract: Speech acts are basic units of linguistic communication which perform actions through words. Certain types of speech acts are especially significant in legislative texts as they go beyond words revealing intentions aiming at shaping the reality of a society. The linguistic theory proposes different types of speech acts. However, this study focuses on the automatic classification of three types for their relevance in the legislative context including: 1) Assertive acts describing events and reality; 2) directive acts setting regulations and, finally, 3) commissive acts indicating commitment to basic rights and principles. For the training and evaluation, a dataset of 1325 statements was manually labeled and further splitted into train and validation sets (80%-20%). Then, the resulting trained classifiers were further evaluated against a test dataset of 250 statements. Different classifier were trained over three types of models: Classical machine learning models, foundational Large Language Models (LLMs) based on "encoders"; namely RoBERTaLex and BERT and finally, generative models based on "decoders", namely GPT3.5 through a 5-shot prompt tuning. The classifier based on encoder LLMs (BERT and RoBERTa) outperformed the rest of models. BERT achieved f1-macro score of 0.85 for all classes and a f1-micro score of 0.87 (BERT) and 0.86 (RoBERTa).
Keywords: Language models, Speech Acts, Computational Pragmatics, Legal Text Processing.

ISSN 1135-5948 DOI 10.26342/2024-72-8

1 Introducción y marco teórico

Los actos de habla son unidades básicas de la comunicación lingüística que permiten realizar acciones a través del lenguaje (Instituto Cervantes, 2023). El primero en introducir esta teoría fue el filósofo británico Austin (1962). Su obra póstuma *Cómo hacer cosas con las palabras* asienta las bases de esta teoría. Siete años más tarde, su discípulo Searle (1969) consolidó la teoría acuñando el término “actos de habla”.

Según esta teoría, en un acto de habla se activan tres dimensiones simultáneamente: la dimensión locutiva, ilocutiva y perlocutiva. La dimensión locutiva se centra en el mensaje refiriéndose al acto físico de emitir el mensaje y su formulación fónica, gramatical y semántica. La ilocutiva se centra en la intención del emisor y la acción que quiere realizar más allá de sus palabras. Por último, la dimensión perlocutiva se centra en el efecto provocado en el receptor por el enunciado.

En la legislación, nos encontramos ante una situación comunicativa concreta donde el emisor es la autoridad competente que expide la ley, i.e. el texto legislativo (entendido como mensaje) y un receptor genérico representando a múltiples posibles destinatarios como ciudadanos, profesionales del derecho, juristas, órganos judiciales, autoridades competentes de aplicar la ley, instituciones públicas o privadas, administraciones del Estado, etc.

En esta situación comunicativa, los actos de habla ilocutivos cobran especial relevancia porque mediante estos actos, el órgano legislador pretende llevar a la realidad acciones que implican organizar las relaciones y establecer las normas.

El marco teórico establece diferentes tipologías de los actos de habla (Austin, 1962; Searle, 1969; Vendler, 1980; Bach y Harnish, 1979). No obstante, el presente estudio se centra en un subconjunto de tres tipos de actos por su relevancia en el contexto legislativo. Los actos asertivos que describen los hechos. Los actos directivos que establecen las normas. Este tipo de actos es el más común en los textos legislativos. Finalmente, los actos compromisorios que reconocen los derechos o muestran un compromiso por velar y proteger estos derechos. Para el resto de los actos fuera del alcance de este estudio se ha contemplado una cuarta categoría “Otros”.

El estudio se estructura en siete secciones incluyendo esta introducción. En la Sección 2 se define el objetivo general, los objetivos específicos y el alcance. La Sección 3 ofrece un análisis del estado de la cuestión. La metodología y la descripción de los datos se describen en la Sección 4. El proceso de anotación se detalla en la Sección 5. Los experimentos para entrenar y adaptar los modelos con los resultados obtenidos se explican en la Sección 6. Finalmente, las conclusiones resumen los principales hallazgos.

2 Objetivo y alcance

2.1 Objetivo

El presente trabajo tiene como objetivo clasificar automáticamente los actos de habla en los textos legislativos distinguiendo entre los actos asertivos, directivos, compromisorios y otros actos que no se incluyen bajo estas tres categorías. Es una clasificación multiclase y multi-etiqueta porque en un solo enunciado se puede expresar más de una intención. Por ejemplo, un enunciado puede indicar una norma (acto directivo) a la vez que reconocer un derecho (un acto compromisorio).

Para alcanzar el objetivo general se plantean los siguientes objetivos específicos:

- Recopilar una muestra representativa de actos de habla a partir de un corpus de textos legislativos.
- Establecer unos criterios para la anotación y clasificación manual de esta muestra.
- Anotar un conjunto de datos de referencia “gold-standard” que sirva tanto para el entrenamiento como para la validación de los diferentes modelos.
- Entrenar clasificadores automáticos basados en diferentes familias de modelos: 1) modelos clásicos de aprendizaje automático (Ej. *RandomForest, OneVsRestClassifier*), 2) modelos fundacionales de lenguaje de tipo “encoder” (*BERT* y *RoBERTaLex*; ambos son modelos pre-entrenados con datos en español) y 3) modelos fundacionales generativos de tipo “decoder” (*GPT 3.5*).
- Evaluar los resultados obtenidos de los diferentes modelos.

Explorar este abanico de modelos permite comparar y evaluar la capacidad de las técnicas actuales en abordar esta tarea y, por tanto, valorar su viabilidad y su posible impacto en soluciones y servicios finales destinados a diferentes tipos de usuarios en el dominio legal. Por ejemplo, la clasificación de actos de habla puede asistir a los profesionales del derecho en extraer la información relevante de los grandes volúmenes de textos legislativos distinguiendo de forma más rápida y eficiente entre los enunciados directivos que establecen una norma de los enunciados compromisorios que reconocen un derecho. Por otro lado, para un ciudadano este tipo de clasificación le puede ayudar a identificar los derechos y las obligaciones en una ley.

Asimismo, esta clasificación puede ser un módulo a integrar en sistemas conversacionales interactivos, sistemas de pregunta-respuesta o en soluciones para la simplificación del texto legislativo, la generación de resúmenes automáticos, etc.

2.2 Alcance

El estudio se centra en la dimensión ilocutiva referente a la intención o la acción que se pretende realizar con las palabras del enunciado. Dentro de esta dimensión ilocutiva, se han seleccionado tres tipos de actos: asertivos, directivos y compromisorios por su relevancia en el texto legislativo siguiendo la clasificación de López-Hernández (2005).

Otros estudios de carácter puramente lingüístico sin un componente computacional han contemplado otros tipos de actos de habla como los cualificatorios o los realizativos/performativos. Los primeros sirven para definir conceptos y los segundos convierten en realidad lo que se dice en el enunciado (López-Hernández, 2005).

Este último tipo es más significativo en los textos jurisprudenciales o sentencias. No se han contemplado en este estudio porque su uso en los textos legislativos es poco significativo y se limita a fórmulas tradicionales para declarar la entrada en vigor de una ley. Por ejemplo, al principio de una ley:

> *Rey de España: A todos los que la presente vieren y entendieren. Sabed: Que las Cortes Generales han aprobado y Yo vengo en sancionar la siguiente ley orgánica*.

Además, los enunciados al final de una ley. Por ejemplo "Por tanto, Mando a todos los españoles, particulares y autoridades, que guarden y hagan guardar esta ley orgánica*"*.

Partiendo de los objetivos y el alcance, el presente estudio plantea tres cuestiones que se detallan a continuación:

- ¿Es viable abordar los actos de habla en el texto legislativo desde una perspectiva computacional mediante el entrenamiento de clasificadores automáticos?
- ¿Qué clasificadores automáticos realizarían mejor la tarea?
- ¿Cuáles son los retos que presenta esta tarea?

3 Estado de la cuestión

El análisis del estado de la cuestión llevado a cabo revela dos enfoques principales: 1) Los estudios de Lingüística y la Filosofía del Derecho que han tratado los actos de habla en el dominio legal desde una perspectiva pragmática y 2) los estudios computacionales que han abordado los actos del habla como un problema de clasificación automática.

Según este análisis, ningún estudio ha abordado los actos del habla en los textos legales como un problema de clasificación automática.

De ahí la aportación del presente estudio donde se retoma la teoría de los actos de habla desde un enfoque pragmático computacional. A la vez, es un enfoque aplicado a los textos legislativos que ofrece una solución automática empleando las técnicas actuales en el ámbito del Procesamiento del Lenguaje Natural (PLN) y la Inteligencia Artificial (IA).

3.1 Actos de habla en textos legales

Numerosos estudios han abordado los actos de habla en los textos legales. No obstante, todos estos estudios han tratado el tema desde una perspectiva puramente lingüística y pragmática o a veces desde la perspectiva de la filosofía del derecho. Ningún estudio de los señalados ha planteado una solución computacional automatizada. Además, varios se centran en el análisis de estos actos en inglés.

En cuanto a tipologías y clasificaciones de los actos de habla, se destaca el estudio pionero de Kurzon (1986) sobre los actos de habla en textos legales. Maccormick y Bankowski (1991) compararon los actos de habla con los actos jurídicos en su trabajo titulado "La teoría de los actos de habla y la teoría de los actos jurídicos".

Siguiendo esta línea, Visconti (2009) agrupó los actos de habla en actos *a praxis* y actos *a poiésis*. Los primeros representan acciones sin pretender cambiar la realidad (Ej. definir conceptos, constatar realidades). El segundo tipo son actos que pretenden cambiar la realidad del mundo a nuestro alrededor (Ej. ordenar, prometer, acusar, etc.). Por último, Durant y Leung (2016) prestaron especial interés a los actos realizativos/performativos.

Por otro lado, se destacan estudios aplicados que analizaron los actos de habla en diferentes tipos de textos legales. En este sentido, Blom y Trosborg (1992) y Trosborg (1995) analizaron cuantitativamente los actos de habla en estatutos y contratos en inglés basándose en un corpus limitado. En un estudio más reciente, Janicki (2018) ha retomado el análisis de los actos de habla en contratos.

Bernal (2007) analizó los actos de habla en las decisiones judiciales. Por otro lado, Mey (2013) abordó los actos de habla en contratos, acuerdos matrimonios, sentencias, etc. desde un enfoque diacrónico. Recientemente, Kone (2020) ha analizado los actos de habla en los tratados de las Naciones Unidas. Finalmente, en su estudio titulado “When does Speech Perform Regulable Action?”, Weston (2022) ha criticado la teoría de los actos de habla aplicándola a la regulación de la libertad de expresión.

En cuanto a los estudios de actos de habla en textos legales en español, se destaca el estudio de López-Hernández (2005) en el que propuso una clasificación de las normas jurídicas como enunciados de actos ilocutivos.

Otros estudios se han centrado en actos de habla concretos o en documentos legales específicos. En esta línea, destacamos los trabajos de Cifuentes-Honrubia (2005; 2006; 2009) sobre la “autorización” como acto de habla. También, los trabajos que enfocaron los actos realizativos/performativos en el discurso legal (Fiorito, 2006**;** Amorebieta y Vera, 2020**)**. Moreu-Carbonell (2020) estudió los actos comunicativos en el lenguaje administrativo destacando la particularidad del lenguaje jurídico español. Por último, Ibáñez-Macías (2021) analizó los actos declaratorios/cualificatorios en el Derecho constitucional.

3.2 Clasificadores automáticos de actos de habla

Desde el enfoque computacional, varios trabajos han desarrollado clasificadores automáticos de actos de habla, pero no en el dominio legal, sino en otros dominios como las redes sociales, los textos de mensajería instantánea, los correos electrónicos, los foros educacionales, etc. Estos estudios se han centrado en la dimensión ilocutiva para una mejor detección de los intentos.

En cuanto a la clasificación de actos en diálogos y mensajería instantánea, se destaca el trabajo pionero de Mast et al. (1996) basado en la clasificación semántica. Twitchell et al. (2004) partieron de la teoría de actos de habla para modelar las conversaciones en mensajería instantánea. Carvalho y Cohen (2005) abordaron los actos de habla en los correos electrónicos para detectar si se tratan de solicitudes o de compromisos. Moldovan, Rus y Graesser (2013) entrenaron un modelo supervisado con árboles de decisiones para las conversaciones en línea. Samei et al. (2014) entrenaron un clasificador de actos de habla para analizar los diálogos de sistemas inteligentes de tutoría (*Intelligent Tutoring Systems*). Arguello y Shaffer (2015) también entrenaron un clasificador para los actos de habla en los foros educativos de los cursos abiertos en línea (*MOOC*).

Recientemente, los actos de habla en *tweets* han sido objeto de varios trabajos. Zhang, Gao y Li (2011) desarrollan una metodología para reconocer “qué hacen los tweets”. Vosoughi y Roy (2016) también analizan los actos de habla en *tweets*. Por último, Saha, Saha y Bhattacharyya (2019) desarrollan un clasificador de actos de habla en *tweets* basado en redes neuronales.

4 Metodología y datos

Para el presente estudio se ha adoptado una metodología empírica y aplicada siguiendo cuatro fases principales:

- Recopilación de los datos.
- Extracción de enunciados.
- Anotación de los enunciados según los actos de habla objeto de este estudio.
- Entrenamiento y evaluación de los clasificadores automáticos.

Los desarrollos se han realizado en el lenguaje de *Python* utilizando las librerías abiertas de:

- *Stanza* (Qi et al., 2020) para el preprocesamiento del texto del corpus y *pandas* (McKinney, 2010) para la estructuración de los conjuntos de datos.
- *Scikit-learn* (Pedregosa et al., 2011) para el entrenamiento y la evaluación de los

modelos clásicos de aprendizaje automático.
- *Transformer*s de *HuggingFace* (Wolf et al., 2020) para el entrenamiento y la evaluación de los modelos fundacionales de tipo "encoder".
- Modelos fundacionales "encoder" pre-entrenados con textos en español: *BERT* [distil-bert] (Cañete et al., 2023) y *RoBERTaLex* (Gutiérrez-Fandiño et al., 2021). Están disponibles con acceso abierto en el repositorio de *HuggingFace*.
- El interfaz conversacional *ChatGPT* (OpenAI, 2023) para el experimento con el modelo fundacional generativo *GPT 3.5*.

4.1 El corpus de textos legislativos

Existen corpus legales como Legal-ES (Samy, Arenas-García y Pérez-Fernández, 2020), el corpus "Spanish Legalese Language Model and Corpora" (Gutiérrez-Fandiño et al., 2021) o el corpus titulado "Spanish monolingual corpus from contents of Spanish State Official Gazette" disponible en *European Language Grid* (2022).

Para este estudio, se ha partido de un subcorpus limitado de textos legislativos, dado que se trata de una tarea concreta y no es necesario contar con un corpus grande. El subcorpus utilizado incluye tres fuentes principales: El Código Civil, la Ley de Protección de Datos Personales y la Ley de Régimen Jurídico de las Administraciones Públicas.

Conjunto de datos	Nº*tokens*	Nº *types*
Código Civil	129251	9732
Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales	40681	3930
Ley de Régimen Jurídico de las Administraciones Públicas y del Procedimiento Administrativo Común	31613	3849
Total	**201545**	**17511**

Tabla 1: Textos legislativos para extraer la muestra de los actos de habla.

A continuación, incluimos ejemplos de los tres tipos de actos de habla: asertivos, directivos, compromisorios junto a la categoría "Otros".

Acto de habla	Ejemplo
Asertivo	Es preciso ahora que el marco que regula el régimen jurídico de las Administraciones Públicas sea objeto de una adaptación normativa expresa que lo configure de forma armónica y concordante con los principios constitucionales.
Directivo	Si el recurso se hubiera interpuesto ante el órgano que dictó el acto impugnado, éste deberá remitirlo al competente en el plazo de diez días, con su informe y con una copia completa y ordenada del expediente.
Compromisorio	Los poderes públicos velarán por los derechos y las necesidades de las personas que hayan padecido daños causados por catástrofes.
Otros	A todos los que la presente vieren y entendieren. Principios de la potestad sancionadora Disposición adicional decimotercera

Tabla 2: Ejemplos de enunciados con los tipos de actos de habla.

4.2 Preprocesamiento

Se ha preprocesado el subcorpus para garantizar su calidad en cuanto a formato, codificación, etc.

En general, los textos procedentes de fuentes oficiales como el Boletín Oficial del Estado (BOE) garantizan una buena calidad en cuanto a formato con poco ruido textual. Por lo tanto, el esfuerzo de la depuración ha sido limitado a ajustar algunos saltos de líneas y casos puntuales de caracteres mal codificados. Se ha utilizado *Stanza* (Qi et al., 2020) para la segmentación de oraciones que han constituido la unidad de análisis para este trabajo.

Cabe destacar que la oración es una unidad de análisis sintáctico-semántico, mientras que los actos de habla es un fenómeno pragmático que pueda coincidir o no con la oración. Esto depende del contexto lingüístico y extralingüístico. Por esto, el fenómeno pragmático puede reflejarse en una o más oraciones. No obstante, se ha decidido basarse en la oración como unidad mínima de análisis en este estudio por razones prácticas de cara al procesamiento computacional.

5 Anotación de los actos de habla

La anotación es una piedra angular para el entrenamiento de modelos de aprendizaje supervisado. La calidad de los datos de entrenamiento y la coherencia de los criterios son factores principales para que el modelo aprenda y generalice mejor. Además de servir como base para el entrenamiento y la evaluación de los modelos de clasificación automática, un conjunto de referencia estándar "gold standard" es en sí un recurso lingüístico de interés.

Para garantizar la calidad del conjunto anotado, el proceso se ha llevado a cabo en tres pasos: 1) extracción de enunciados que representan los tipos de actos de habla en cuestión, 2) establecimiento de los criterios y la anotación manual y 3) validación de la anotación. A continuación, describimos brevemente cada paso.

5.1 La extracción de los actos de habla

Con el corpus preprocesado y segmentado en oraciones, se procede a la extracción de enunciados para anotar los actos de habla.

Existen algunos retos a la hora de afrontar la tarea de extracción de enunciados que representan los tres tipos de actos de habla objeto de este estudio.

Primero, no todas las oraciones contienen los tipos de actos de interés (asertivos, directivos y compromisorios). Segundo, los actos de habla son fenómenos pragmáticos a nivel del enunciado que no siempre coincide con el límite de la oración. Por lo tanto, la interpretación del acto de habla a nivel de oración requiere de un conocimiento del contexto extralingüístico por parte del anotador. Tercero, la ocurrencia de los tres tipos está desequilibrada.

Por ejemplo, identificar actos de habla directivos es más fácil porque son bastante frecuentes, mientras que los actos compromisorios son menos frecuentes y habría que examinar muchos enunciados para localizar ejemplos de actos compromisorios.

Por estos motivos, se ha incluido una cuarta categoría "Otros" en la que se contemplan los enunciados que no se clasifican bajo ningún tipo de los tres tipos de actos señalados. Asimismo, se ha recurrido a algunas estrategias para aumentar los datos de los actos compromisorios para tener una cantidad suficiente a la hora de entrenar y evaluar los modelos.

Los datos anotados se han extraído en 3 conjuntos sumando un total de 1375 oraciones. Los dos primeros conjuntos consisten en muestras aleatorias.

Para el tercer conjunto, se han extraído muestras dirigidas de enunciados principalmente de tipo compromisorio. Se ha optado por esta estrategia para equilibrar esta categoría de cara al entrenamiento y la evaluación, ya que esta clase se ha quedado infrarrepresentada en las muestras aleatorias.

En la búsqueda dirigida solo aplicada al caso de los actos compromisorios, se han extraído como posibles candidatos enunciados que contengan lemas como "garantizar", "velar", "derecho". Sin embargo, es importante señalar que no todas las menciones a derechos o garantías son necesariamente actos de habla compromisorios. Existen varios casos de ejemplos en los que no se aplican los criterios de actos compromisorios.

Por este motivo y para evitar generalizaciones imprecisas, es necesario tener en cuenta estas estrategias.

Primero, incluir ejemplos negativos donde aparecen formas de "velar", "garantizar" o "derecho" sin que se trate de un acto compromisorio. Por ejemplo, en el enunciado "La Ley recoge esta concepción constitucional [...] y fija las garantías mínimas de los ciudadanos [...]", no se trata de un acto compromisorio, sino de un acto asertivo que describe una realidad sobre la ley.

Segundo, resolver estos casos ambiguos mediante el proceso de anotación manual.

Tercero, incluir una clase "Otros" para aquellos casos en los que no se cumplen los criterios de ninguno de los tres tipos (asertivos, compromisorios o directivos).

De esta forma, se garantiza una mejor representatividad de la muestra.

La Figura 1 muestra la distribución en el conjunto completo de entrenamiento y validación (1075 enunciados) después de equilibrar la categoría de actos compromisorios.

Del total de 1375, se han descartado aproximadamente 50 oraciones porque presentaban problemas resultantes del proceso de segmentación. Por ejemplo, algunas eran solamente un dígito o un fragmento mal segmentado de una referencia a una ley por la presencia de barras o guiones, etc.

Figura 1: Distribución de actos de habla en el conjunto de entrenamiento y validación (1075 enunciados).

De ahí, el total anotado es 1325 enunciados con cuatro categorías: actos asertivos, directivos, compromisorios u otros. Para el proceso de entrenamiento y validación se ha utilizado un conjunto de 1075 enunciados. Para la prueba de los modelos ya entrenados, se ha utilizado un conjunto de 250 enunciados.

5.2 Criterios de anotación de los actos de habla

Para la anotación manual, se han establecido unos criterios para distinguir las tres categorías de actos de habla ilocutivos (asertivos, directivos y compromisorios) donde la intención es la base principal para la clasificación. Los enunciados que no se clasifican bajo ninguna de las tres categorías, se clasifican con la etiqueta "Otros". Los criterios establecidos incluyen lo siguiente:

- **Asertivos**. Se anotan como actos asertivos las oraciones que definen una realidad o relatan unos hechos. Este tipo de actos es común en las disposiciones de una ley. En un texto de ley, las disposiciones suelen estar al principio del texto legislativo y suelen reflejar el contexto y los motivos por los cuales, se ha elaborado esta ley.
- **Directivos**. Se anotan como actos directivos los enunciados cuya interpretación indica establecer unas normas, regular ciertas relaciones o definir competencias. El acto directivo puede ser directo empleando modalidades deónticas como "deber", "haber de", "estar prohibido". También el acto directivo puede ser indirecto cuando se describen unos escenarios hipotéticos como "El responsable entregará en el plazo de un mes […]", "la autoridad competente convocará", etc.
- **Compromisorios.** Se anotan como actos compromisorios aquellos enunciados que reconozcan los derechos y se comprometan a velar por ellos de forma directa o indirecta. En los casos directos, se indica de forma explícita como por ejemplo "velar por" o "respetando el derecho". En los casos indirectos, el texto de la ley compromete a otros órganos o agentes indicando que es parte de su responsabilidad. Por ejemplo: "Es responsabilidad de […] velar por el cumplimiento del derecho de igualdad".
- **Otros**. En los casos donde no se cumple ninguno de los criterios establecidos, se anotan como "Otros".
- En los casos ambiguos, se recurre al contexto original de dónde se ha extraído la muestra para aclarar dudas. Si persiste la ambigüedad, se da prioridad al acto directivo teniendo en cuenta su relevancia en el texto legislativo.
- **Casos multi-etiqueta**. Los enunciados que realizan dos acciones se anotan con los actos correspondientes (multi-etiqueta). Es decir, si en un enunciado, se describe una realidad y luego se establece una norma en el mismo enunciado, se anotan las dos categorías: el acto asertivo y el acto directivo.
- Del mismo modo, se anotan con las categorías directiva y compromisoria los enunciados donde a la vez se establece una norma y se reconoce el compromiso por un derecho. Ejemplo de ello es este enunciado:

  Cada Administración pública establecerá los días y el horario en que deban permanecer abiertos sus registros, garantizando el derecho de los ciudadanos a la presentación de documentos.

En total, los enunciados con multi-etiquetas representan sólo el 3,4% del conjunto de entrenamiento y validación.

5.3 Validación de la anotación

Una vez concluido el proceso de anotación, se procede a la validación de las muestras anotadas, sobre todo, los casos ambiguos para tomar una decisión acerca de su clasificación.

En un escenario óptimo, la anotación debería llevarse a cabo por varios anotadores. Luego, en el proceso de validación se calculan los acuerdos entre anotadores para proceder a la armonización.

No obstante, en este proyecto, no se ha contado con recursos para tener un equipo de anotación. La mayoría del proceso se ha realizado por un solo anotador, un lingüista experto. Por eso, la validación consistía en revisar los casos ambiguos y asegurarse del cumplimiento de los criterios de una forma coherente.

6 Entrenamiento

El total de enunciados anotados y validados (1325) se han dividido en: 1075 enunciados para el entrenamiento y la validación, por un lado y 250 enunciados como un conjunto de prueba final para evaluar los modelos entrenados. Los datos anotados y los modelos entrenados se encuentran disponibles en el repositorio de *Github*.[1]

El primer conjunto de 1075 enunciados se ha dividido de forma aleatoria en dos subconjuntos de entrenamiento (80%) y validación (20%) asegurándose de que: 1) estén representadas todas las categorías; 2) los conjuntos de entrenamiento y validación sean iguales en todos los experimentos para garantizar la comparabilidad de los resultados

6.1 Modelos clásicos de aprendizaje automático

Se han entrenado clasificadores basados en modelos clásicos disponibles en la librería de *scikit-learn* como *RandomForest* u *OneVsRestClassifier* en combinación con vectorizadores como *Tf-idf* o *HashingVectorizer*. Estos modelos tienen la ventaja de ser eficientes y requieren pocos recursos de cómputo para entrenarse. El inconveniente es que dependen de las palabras, i.e. el léxico empleado y, por lo tanto, tienen limitaciones a la hora de generalizar.

Los mejores resultados se han obtenido entrenando un clasificador *OneVsRestClassifier* con un vectorizador *Tf-idf* obteniendo un *f1-macro* de 0,74 y un *f1-micro* de 0,77 en el conjunto de prueba.

Los actos directivos y compromisorios obtuvieron mejores resultados que los actos asertivos. En nuestra opinión, esto se debe a: 1) los actos directivos son los más frecuentes; 2) los actos asertivos, en cambio, son más ambiguos y más variados sin rasgos distintivos lingüísticamente y 3) los actos compromisorios, sí, se caracterizan por el empleo de un léxico concreto o patrones más específicos.

La Tabla 3 muestra los resultados obtenidos por este tipo de modelos. No obstante, para una evaluación más rigurosa, se ha realizado una validación cruzada con 5 y 10 pliegues. Con la validación cruzada, los resultados bajan significativamente demostrando la limitación de estos modelos.

Modelos clásicos			
Conjunto: Validación			
	Precisión	***Recall***	***f1-score***
Asertivo	0,87	0,38	0,53
Directivo	0,81	**0,87**	**0,84**
Compromisorio	**0,91**	0,77	0,83
Otros	0,76	0,74	0,75
f1-score micro	0,77		
f1-score macro	0,74		
Validación cruzada			
	CV = 5	**CV =10**	
f1-micro	0,48	0,53	
f1-macro	0,23	0,22	
Accuracy (mean)	0,36	0,42	
Conjunto: Prueba (250 enunciados)			
Asertivo	0,61	0,19	0,29
Directivo	**0,91**	**0,72**	**0,84**
Compromisorio	0,70	0,47	0,56
Otros	0,69	0,62	0,65
f1-score micro	0,70		
f1-score macro	0,58		

Tabla 3: Resultados del modelo clásico OneVsRestClassifier.

[1] https://github.com/dosamy/SpeechActs-Legislative-Spanish

6.2 Modelos fundacionales basados en *encoder*

Los grandes modelos fundacionales han supuesto un salto cualitativo y cuantitativo en el panorama del PLN. Los modelos de tipo “encoder” han demostrado una mejora sustancial en las tareas de comprensión del lenguaje (*Natural Language Understanding* - *NLU*).

La hipótesis inicial considera que basarse en este tipo de modelos “encoder” puede suponer un avance significativo respecto a los modelos clásicos de aprendizaje automático abordados en la sección anterior.

Se han realizado varias rondas de entrenamiento para la adaptación *fine-tuning* de dos modelos fundacionales de tipo “encoder”; ambos pre-entrenados y adaptados a la lengua española: *RoBERTaLex* y *BERT*. El primero es pre-entrenado con textos del dominio legal, lo cual se supone que pueda obtener mejores resultados en nuestra tarea. El segundo es pre-entrenado con textos del dominio general.

Ambos modelos se han entrenado mediante la librería de Transformers en 10 épocas aplicando los mismos parámetros (*Learning rate=2e-5, batch_size=8, weight*) en un ordenador con CPU.

Las Tablas 4 y 5 muestran los resultados obtenidos por *RoBERTaLex* y *Distil-BERT*.

Modelos fundacionales *encoder*: RoBERTaLex			
Conjunto: Validación			
	Precisión	***Recall***	***f1-score***
Asertivo	0,75	0,73	0,74
Directivo	0,89	**0,92**	0,90
Compromisorio	**0,92**	**0,92**	**0,92**
Otros	0,86	0,88	0,87
f1-score micro	0,86		
f1-score macro	0,86		
Conjunto: Prueba (250 enunciados)			
Asertivo	0,59	0,47	0,52
Directivo	**0,95**	0,83	**0,88**
Compromisorio	0,82	0,60	0,69
Otros	0,65	**0,96**	0,78
f1-score micro	0,78		
f1-score macro	0,72		

Tabla 4: Resultados del modelo *RoBERTaLex*.

Los resultados de *RoBERTaLex* como un modelo fundacional pre-entrenado en el dominio legal superan significativamente los modelos clásicos alcanzando un *f1-score micro* de 0,86 y un *f1-score macro* de 0,86 en el conjunto de validación, y 0,78 y 0,72 respectivamente en el conjunto de prueba.

Por otro lado, se ha replicado el mismo proceso de adaptación “fine-tuning” con el modelo *BERT* pre-entrenado para la lengua española.

Modelos fundacionales *encoder*: BERT			
Conjunto: Validación			
	Precisión	***Recall***	***f1-score***
Asertivo	0,71	0,69	0,70
Directivo	0,90	**0,94**	0,92
Compromisorio	**1,00**	0,88	**0,94**
Otros	0,85	0,85	0,85
f1-score micro	0,87		
f1-score macro	0,85		
Conjunto: Prueba (250 enunciados)			
Asertivo	0,55	0,47	0,50
Directivo	**0,96**	**0,83**	**0,89**
Compromisorio	0,92	0,80	0,86
Otros	0, 64	0,89	0,74
f1-score micro	0,78		
f1-score macro	0,75		

Tabla 5: Resultados del modelo *BERT*.

Las diferencias entre los modelos *RoBERTaLex* y *BERT* son sutiles. Los resultados obtenidos por *BERT* son mejores, pero es una diferencia mínima. En este sentido, es un resultado inesperado, ya que la hipótesis inicial suponía que el modelo de *RoBERTaLex* obtendría los mejores resultados al estar adaptado al dominio legal en español. No obstante, esta ventaja no ha supuesto una diferencia significativa en la tarea en cuestión.

Por otro lado, es evidente que la tarea sigue suponiendo cierta dificultad, sobre todo en los actos asertivos por su gran variedad y por su difícil distinción de otros tipos de actos.

Pese a ello, cabe destacar que los modelos pre-entrenados han supuesto una mejora significativa respecto a los modelos clásicos, sobre todo en la cobertura.

No obstante, los resultados empeoran en el conjunto de prueba, lo cual podría indicar que los modelos no generalizan lo suficientemente bien y requieren más datos de entrenamiento.

6.3 Pruebas con modelos fundacionales generativos basados en *decoder*

Por último y dado el gran éxito que están teniendo los modelos fundacionales generativos, se ha realizado una prueba con el modelo *GPT3.5* siguiendo una instrucción con 5 ejemplos (*5-shot prompting*) a través del interfaz conversacional ChatGPT.

Estos modelos están entrenados para realizar varias tareas de forma genérica, por lo tanto, utilizarlos en dominios y tareas específicas requiere una adaptación o bien mediante un proceso de "fine-tuning" del modelo o bien a través de una serie de instrucciones "prompting". La adaptación "fine-tuning" es costosa porque supone una modificación de los parámetros del pre-entrenamiento y requiere de procesadores gráficos *(GPU*) y grandes capacidades de cómputo. La adaptación mediante instrucciones es más viable, pero ofrece resultados poco precisos.

Para la prueba realizada, se ha creado una instrucción con 5 ejemplos y se ha introducido esta instrucción a través del interfaz de diálogo *ChatGPT*. En la instrucción, se le ha indicado al modelo el objetivo de la tarea y las clases a tener en cuenta[2].

Se ha introducido el conjunto de prueba formado por 250 enunciados y se han calculado las métricas de los resultados obtenidos por *GPT 3.5*. Con 5 ejemplos y sin un proceso de adaptación es difícil que un modelo generativo realice una tarea tan específica en un dominio tan concreto como el legislativo. Pese a ello, en los actos directivos, la categoría más frecuente, el modelo *GPT 3.5* ha alcanzado un *f1-score* de 0.70.

***ChatGPT 3.5* - Conjunto: Prueba**	
f1-score micro	0,56
f1-score macro	0,48

Tabla 6: Resultados de *ChatGPT 3.5* (*5-shots*).

7 Conclusiones

El problema de clasificación al que nos enfrentamos en este estudio es un problema multi-etiqueta y multiclase de una dimensión de cuatro clases. Es un problema complejo a nivel lingüístico por la ambigüedad de algunos actos.

Los resultados obtenidos con modelos que representan el estado del arte son bastante satisfactorios (0,87 de f-*score*) aunque dejan un margen de mejora.

A nuestro modo de ver, los retos que supone la tarea se resumen en lo siguiente:

- Es una tarea que se basa principalmente en un criterio pragmático, i.e. la intención del emisor, que va más allá del texto. A pesar de que el emisor en el texto legislativo es una autoridad competente que suele seguir un estilo de lenguaje normalizado, el hecho de que se trate de la interpretación de la intención puede ser un reto, incluso para la anotación humana.
- La segmentación y la variabilidad de la extensión de los enunciados es otro factor que puede suponer un reto porque a veces se trata de enunciados cortos y otras veces se trata de enunciados extensos y complejos.
- Los enunciados que se clasifican con más de una etiqueta es otro reto por su complejidad conceptual y porque solo representan un porcentaje limitado que no alcanza el 5%.
- A modo de conclusión, los resultados obtenidos son satisfactorios, pero para una mejora, es necesario ampliar el conjunto de datos de entrenamiento para incluir más casos de multi-etiqueta, de actos asertivos y de actos compromisorios.

Por otro lado, una línea de mejora es la adaptación de un modelo generativo como *Llama 2* o *GPT 4.0*. No obstante, es un proceso que requiere de mayores recursos de cómputo.

Por último, con este estudio se abre una línea de investigación para evaluar la utilidad de incluir este tipo de clasificadores en algunos servicios y productos. Sistemas de preguntas-respuestas, generación de resúmenes, simplificación de textos legislativos son ejemplos de soluciones que pueden beneficiarse de la clasificación como un paso relevante para mejorar el resultado final.

Bibliografía

Amorebieta y Vera, J. 2020. Las sentencias de la Corte Penal Internacional: análisis preliminar del uso de los verbos performativos. *Quintú*

[2] La instrucción está disponible en el repositorio de *Github*.

Quimün Revista De lingüística. Número 4, páginas 1-13.

Arguello, J. y K. Shaffer. 2015. Predicting speech acts in MOOC forum posts. En *Proceedings of the International AAAI Conference on Web and Social Media, ICWSM,* 9:1, páginas. 2-11, Oxford (Reino Unido).

Austin, J. L. 1962. *Cómo hacer cosas con palabras*. Paidós, Barcelona.

Bach, K. y R.M. Harnish. 1979. *Communication and Speech Acts*. MIT Press, Massachusetts.

Bernal, C. L. 2007. A Speech Act Analysis of Judicial Decisions. *European Journal of Legal Studies*. 1:2, páginas 1-24.

Blom, B., y A. Trosborg. 1992. An Analysis of Regulative Speech Acts in English Contracts - Qualitative and Quantitative Methods. *HERMES - Journal of Language and Communication in Business*. 5:9, páginas 83–111.

Cañete, J., G. Chaperon., R. Fuentes, J. Ho, H. Kang y J. Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. En *Proceedings of Practical ML for Developing Countries: learning under limited/low resource scenarios Workshop at the Eighth International Conference on Learning Representations, PML4DC@ICLR 2020,* páginas 1-9, Addis Ababa (Etiopía). Modelo disponible en: https://huggingface.co/dccuchile/distilbert-base-spanish-uncased

Cao, D. 2007. Legal Speech Acts as Intersubjective Communicative Action. En Wagner, A., W. Werner y D. Cao (Eds.), *Interpretation, Law and the Construction of Meaning*, páginas 65–82. Springer, Dordrecht.

Carvalho, V. R. y W. Cohen. 2005. On the collective classification of email speech acts. En *Proceedings of the 28th Annual ACM/SIGIR International Conference on Research and Development in Information Retrieval*, páginas. 345-352, Salvador (Brasil).

Cifuentes-Honrubia, J.L. 2006. *Los actos de habla*. Liceus, Madrid.

Cifuentes-Honrubia, J.L. 2009. Autorización: actos de habla e interpretación jurídica. En Jiménez Ruiz, J.L. y L. Timofeeva (Eds.), *Estudios de lingüística: Investigaciones lingüísticas en el siglo XXI, ELUA*, páginas 45-61, Universidad de Alicante.

Durant, A. y H. Leung. 2016. *Language and Law*. Routledge, Oxon.

European Language Grid. 2022. *Spanish monolingual corpus from contents of Spanish State Official Gazette*. Dataset (Text corpus). Disponible en: https://live.european-language-grid.eu/catalogue/cpid/mxaKn3VZmpuKCfGhEdSPUX/

Fiorito, L. 2006. On Performatives in Legal Discourse. *Metalogicon*, 19:2, páginas 101-112.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, A. Gonzalez-Agirre y M. Villegas. 2021. *Spanish legalese language model and corpora*. arXiv. DOI: arXiv:2110.12201. Modelo disponible en: https://huggingface.co/PlanTL-GOB-ES/RoBERTale

Ibáñez-Macías, A. 2021. Las declaraciones como actos de habla en el Derecho constitucional. *Revista Telemática de Filosofía del Derecho*. (24), páginas 3-36.

Janicki, E. A. 2018. Contracts as speech acts: Bringing Jakobson to the conversation. *GeorgeTown Law Journal*. 107:2018, páginas 201-233.

Kone, N. 2020. Speech Acts in UN Treaties: A Pragmatic Perspective. *Open Journal of Modern Linguistics*. 10:6, páginas 813-827.

Kurzon, D. 1986. *It is Hereby Performed: Explorations in Legal Speech Acts*. John Benjamins Publishing, Amsterdam.

López-Hernández, J. 2005. Clasificación de las normas jurídicas como enunciados de actos ilocutivos. *Anuario de derechos humanos*. (6), páginas. 455-510.

López-Hernández, J. 2005. Las normas jurídicas como actos ilocutivos: Conceptos y clases. *Cuadernos Electrónicos de Filosofía del Derecho*. (11), páginas 1-48.

MacCormick, N. y Z. Bankowski. 1991. La teoría de los actos de habla y la teoría de los actos jurídicos. *Anuario de Filosofía del derecho*. (8), páginas 219-238.

Mast, M., H. Niemann, E. Nöth y E.G. Schukat-Talamazzini. 1996. Automatic classification of dialog acts with Semantic Classification Trees and Polygrams. En: Wermter, S., E. Riloff, y G. Scheler. (Eds.), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing. International Joint Conference on Artificial Intelligence, IJCAI 1995*. Lecture Notes in Computer Science. (1040), páginas 217-229, Springer, Berlin, Heidelberg.

McKinney, 2010. Data structures for statistical computing in python. En *Proceedings of the 9th Python Science Conferenc*e. (445), páginas 56-61.

Mey, J. L. 2013. Atos de fala e a Lei. *Cadernos De Linguagem E Sociedade*. 4:2, páginas 11–27.

Moldovan, C., V. Rus y A.C. Graesser. 2011. Automated Speech Act Classification For Online Chat. En *Proceedings of the 22nd Midwest AI and Cognitive Science Conference, MAICS*, páginas 23-29, Ohio (Estados Unidos).

Moreu Carbonell, E. 2020. Nuestro lenguaje: el giro lingüístico del derecho. *Revista de Derecho Público: Teoría y Método.* (1), páginas 313-362.

OpenAI 2023. ChatGPT [Large language model]. Disponible en: https://chat.openai.com/chat

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel y E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*. (12), páginas 2825-2830.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton. y C.D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, *ACL* 2020, páginas 101–108.

Robles-Morchón, G. 2009. *Comunicación, Lenguaje y Derecho: algunas ideas básicas de la teoría comunicacional del Derecho*. Real Academia de Ciencias Morales y Políticas, Madrid.

Saha, T., S. Saha y P. Bhattacharyya. 2019. Tweet act classification: A deep learning based classifier for recognizing speech acts in Twitter. En *International Joint Conference on Neural Networks, IJCNN*, páginas 1-8, Budapest (Hungary).

Samei, B., H. Li, F. Keshtkar, V. Rus y A.C. Graesser. 2014. Context-based speech act classification in intelligent tutoring systems. En *International conference on Intelligent Tutoring Systems*, *ITS,* páginas 236-241, Honolulu Hi (Estados Unidos).

Samy, D., J. Arenas-García y D. Pérez-Fernández. 2020. Legal-ES: A Set of Large-Scale Resources for Spanish Legal Text Processing. En Samy, D. J. Arenas-García y D. Pérez-Fernándes (Eds.) *Proceedings of Workshop on Language Technologies in Government and Public Administration, LT4Gov@LREC-2020*, páginas 32-36, Marseille (Francia).

Searle, J. 1969. *Speech Acts.* Cambridge University Press, Cambridge.

Trosborg, A. 1995. Statutes and contracts: An analysis of legal speech acts in the English language of the law. *Journal of Pragmatics*. 23:1, páginas 31-53.

Twitchell, D.P, M. Adkins, J.F. Nunamaker y J.K. Burgoon. 2004. Using Speech Act Theory to Model Conversations for Automated Classification and Retrieval. En *Proceedings of the 9th International Working Conference on the Language-Action Perspective on Communication Modelling*, *LAP*, páginas 121-130, Tilburg (Holanda).

Vendler, Z. 1980. Telling the facts. En Searle, J.R., F. Kiefer y M. Bierwisch (Eds.), *Speech Act Theory and Pragmatics. Texts and Studies in Linguistics and Philosophy*, páginas 273-290. Springer, Dordrecht.

Visconti, J. 2009. Speech acts in legal language: introduction. *Journal of Pragmatics*. 41:3, páginas 393-400.

Vosoughi, S., y D. Roy. 2016. Tweet Acts: A Speech Act Classifier for Twitter. En *Proceedings of the International Conference on Web and Social Media*, *AAAI.* 10:1, páginas 711-714, Cologne (Alemania).

Weston, J. 2022. When Does Speech Perform Regulable Action? — A Critique of Speech Act Theory's Application to Free Speech

Regulation. *International Journal of Language and Law*. (11), páginas 78-97.

Wolf, T., D. Lysandre, S. Víctor, J. Chaumon, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. Von Platen, C. Ma, Y. Jernite, J. Plu,C. Xu, T. Le Scao, S. Gugger y M. Drame. 2020. Transformers: State-of-the-art natural language processing. En *Proceedings of the 2020 conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP,* páginas 38-45.

Yin, H. y Y. Chen. 2020. Speech act in diplomacy: How China makes commitments in diplomatic press conference. *International Journal of Languages, Literature and Linguistics*. 6:4, páginas 170-176.

Zhang, R., D. Gao y W. Li. 2011. What Are Tweeters Doing: Recognizing Speech Acts in Twitter. En *Analyzing Microtext: Papers from the 2011 AAAI Workshop*, páginas 86-91, San Francisco (Estados Unidos).

Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 123-136 recibido 15-12-2023 revisado 29-01-2024 aceptado 03-02-2024

Analysing the Problem of Automatic Evaluation of Language Generation Systems

Analizando el Problema de la Evaluación Automática de los Sistemas de Generación de Lenguaje

Iván Martínez-Murillo, Paloma Moreda, Elena Lloret
Department of Language and Computing System
University of Alicante
{ivan.martinezmurillo, moreda, elena.lloret}@ua.es

Abstract: Automatic text evaluation metrics are widely used to measure the performance of a Natural Language Generation (NLG) system. However, these metrics have several limitations. This article empirically analyses the problem with current evaluation metrics, such as their lack of ability to measure the semantic quality of a text or their high dependence on the texts they are compared against. Additionally, traditional NLG systems are compared against more recent systems based on neural networks. Finally, an experiment with GPT-4 is proposed to determine if it is a reliable source for evaluating the validity of a text. From the results obtained, it can be concluded that with the current automatic metrics, the improvement of neural systems compared to traditional ones is not so significant. On the other hand, if we analyse the qualitative aspects of the texts generated, this improvement is reflected.
Keywords: Natural Language Generation, evaluation metrics, NLG architectures, language models.

Resumen: Las métricas automáticas de evaluación de texto se utilizan ampliamente para medir el rendimiento de un sistema de Generación de Lenguaje Natural (GLN). Sin embargo, estas métricas tienen varias limitaciones. Este artículo propone un estudio empírico donde se analiza el problema que tienen las métricas de evaluación actuales, como la falta capacidad que tienen estos sistemas de medir la calidad semántica de un texto, o la alta dependencia que tienen estas métricas sobre los textos contra los que se comparan. Además, se comparan sistemas de GLN tradicionales contra sistemas más actuales basados en redes neuronales. Finalmente, se propone una experimentación con GPT-4 para determinar si es una fuente fiable para evaluar la calidad de un texto. A partir de los resultados obtenidos, se puede concluir que con las métricas automáticas actuales la mejora de los sistemas neuronales frente a los tradicionales no es tan significativa. En cambio, si se analizan los aspectos cualitativos de los textos generados, si que se refleja esa mejora.
Palabras clave: Generación de Lenguaje Natural, métricas de evaluación, arquitecturas de generación, modelos de lenguaje.

1 Introduction

Natural Language Generation (NLG) is a sub-field within the Natural Language Processing (NLP) field that has rapidly evolved in recent years, attracting the interest of the scientific community (Ji et al., 2023). Its evolution and the subsequent advancements have provoked a change in the NLG architectures paradigm, going from architectures that split the generation of language into different sub-tasks to architectures that perform all the generation in just a single task (Gatt and Krahmer, 2018). Two breakthroughs can be highlighted to understand this paradigm shift. On the one hand, the development of deep learning methods improved the state of the art of NLG, generating more coherent and natural text by capturing complex language patterns and context (Gatt and Krahmer, 2018). On the other hand, the proposal of the Transformers architecture (Vaswani et al., 2017) increased considerably the performance of the models, thanks to its self-attention mechanism.

However, these developments have not been reflected in the automatic evaluation of texts. Evaluating the quality of NLG sys-

ISSN 1135-5948 DOI 10.26342/2024-72-9

tems' output still remains a challenge (Dong et al., 2023). There is a lack of standard automated evaluation metrics, as the traditional metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), OR SPICE (Anderson et al., 2016) tend not to correlate well with human judgements (Khapra and Sai, 2021). Most of these metrics rely on comparing a candidate sentence to one or several target sentences, based on a feature overlap (i.e., words or fragments). Thus, the resulting score for the candidate sentence largely depends on how it aligns with the reference sentence (Tang et al., 2023).

This issue can be accentuated in some tasks where the output can vary in content and does not need to follow a predefined structure. Therefore, these texts can also be valid, although quite different from the target sentences. For instance, consider the CommonGen task (Lin et al., 2020), where the model is given a tuple of words and a sentence that incorporates all those words must be generated. Then, that generated sentence is compared against a set of target sentences. In this task, there can be various candidate sentences that can be valid, although they differ from the target sentences. Consequently, when relying on classical evaluation metrics, a model that produces such different sentences might be penalised, even though the generated sentence is both syntactically and semantically correct. Figure 1 shows an example of the CommonGen task, for different models.

To address this, researchers have explored the hypothesis that Large Language Models (LLMs) can exhibit strong correlations with human judgements, being a more suitable approach for evaluating text (Tang et al., 2023). In this line, novel metrics based on LLMs have been proposed, including BARTScore (Yuan, Neubig, and Liu, 2021) and GPTScore (Fu et al., 2023). However, despite efforts to find more suitable metrics, the most popular metrics to evaluate text are still the traditional evaluation metrics.

Given the issues that may arise when evaluating NLG models, this paper aims to empirically analyse and compare different evaluation metrics in the context of a relatively recent task, commonsense generation. Comonsense generation is the task of reasoning about the commonsense while generating coherent text. This task can favour the generation of texts that are not only correct but also diverse, as including commonsense in the generation can produce multiple valid outputs while being semantically different (Yu et al., 2022). Specifically, this work focuses on the CommonGen shared task, previously mentioned. Sentences will be generated using different NLG architectures (traditional and recent) and evaluated based on different metrics during the experimentation. Our goal is to address the following three research questions: (1) Which type of NLG architectures perform best for commonsense generation? (2) How good are the generated sentences of the best-performing system from a qualitative perspective? (3) To what extent can LLMs-based tools, such as ChatGPT, be a good alternative or complement to automatic evaluation metrics?

2 Related Work

2.1 NLG Architectures

The first NLG works date from the decade of 1970 (McDonald, 2010). Since that time, numerous NLG approaches have been proposed, and they can be broadly categorised into three distinct groups according to their architecture (Gatt and Krahmer, 2018):

- *Modular architectures:* This group of approaches considers language generation as a process of three well-differentiated stages. (1) Macroplanning, which includes all the tasks that select what information should be included in the generated text, (2) Microplanning, which includes all the tasks related to the parsing of that information, selecting how to say the information selected previously and (3) Realisation, which perform the tasks of generating the selected information by applying the correct syntactical and grammatical rules. Reiter proposed the standard architecture of this group (Reiter, 1994), consisting of a sequential pipeline of those previously mentioned stages. Other examples of this architecture can be found in (Mann and Moore, 1981), (Hovy, 1987), (Levelt, 1989), and (Nirenburg, Lesser, and Nyberg, 1989).

- *Planning perspectives:* This group of approaches still considered language generation as a process of different stages, but they needed a smaller number of tasks

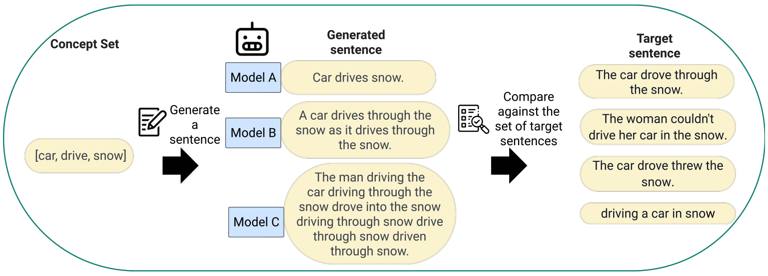


Figure 1: Ilustrative example of the CommonGen task.

to perform the generation. Some examples of approaches within this group are: (Appelt, 1985), (Bateman, 1997),(Koller and Stone, 2007), (Rieser and Lemon, 2009), (Nakatsu and White, 2010) and (Lemon, 2011).

- *Global approaches:* This group is the most predominant in recent years. They do not distinguish between tasks, performing the generation process in one step. An important architecture in this group is the Transformers proposed by (Vaswani et al., 2017), which significantly improved the performance of the NLP field with the concept of self-attention. Other research works using alternative architectures to Transformers are: Graph Neural Networks (Scarselli et al., 2008), Generative Adversarial Nets (Mirza et al., 2014), Recurrent Neural Networks (Sutskever, Vinyals, and Le, 2014), and Memory Networks (Sukhbaatar et al., 2015).

2.2 Commonsense Generation

The task we want to focus on in this empirical analysis is the commonsense generation; therefore, we will briefly describe and contextualise it. LLMs tend to base their predictions on the likelihood of relationships between words, so they lack a fundamental characteristic in human communication, the commonsense. Commonsense knowledge refers to the information that is widely accepted in everyday life (Bhargava and Ng, 2022). Integrating commonsense knowledge in the human language has been recognised as an important and challenging task in the NLG field (Wang et al., 2021), as there is a need to enhance the capability of NLG systems of integrating it in their outputs. Therefore, some shared tasks have been proposed to advance the state of the art of the commonsense generation. In the *Avicenna* (Aghahadi and Talebpour, 2022) task, a model is provided with two premises containing a syllogistic relation. The objective is to generate a conclusion that completes that relation. Integrating commonsense in keyword-to-text task have also been studied. For instance, *SituatedGen* task (Zhang and Wan, 2024) involves generating a pair of contrastive sentences, given a group of concepts that includes temporal or geographical entities. *CommonGen* (Lin et al., 2020) and *C^2Gen* (Carlsson et al., 2022) tasks consist of generating a logical sentence describing an everyday scenario given a set of words. Additionally, the *C^2Gen* task also gives as input a context to which the generated text has to adhere.

2.3 NLG Evaluation Metrics

To evaluate NLG systems is essential to be able to compare and monitor the advancements in the field. While human evaluators would be the most accurate, this type of evaluation is usually impractical due to the temporal and economic cost it involves. Consequently, automatic evaluation metrics can be a viable alternative (Khapra and Sai, 2021). In that line, rule-based metrics have been used for many years, and widely adopted for evaluating different NLG tasks (Sai, Mohankumar, and Khapra, 2022). These metrics tend to evaluate the quality of a text by comparing it against a reference text based on features such as words, characters or embedding. Word-based metrics are the most employed, measuring the word overlapping between a candidate sentence and a reference sentence. This group include metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam, Lawrence Zitnick, and Parikh, 2015), or SPICE (Anderson et al., 2016). Character-based metrics tend to align better with morphologically rich languages comparing the characters of a text instead of the words. Extended Edit Distance (Stanchev, Wang, and Ney, 2019) or chrF (Popović, 2015) are metrics from this group. Finally, embedding-based metrics capture better the semantic similarity in some contexts, by comparing the similarity between the embeddings. Some metrics within this group are: MEANT (Lo, Tumuluru, and Wu, 2012), BERTscore (Zhang et al., 2020), and Word Mover-Distance (Kusner et al., 2015). Other recent types of approaches have studied to use of LLMs to evaluate semantic similarity, usually with a higher correlation with human evaluation from a deep semantic perspective. Some works have explored the accuracy of using ChatGPT as an NLG evaluator (Wang et al., 2023). Inside this group of approaches some metrics have been proposed, e.g. BARTScore (Yuan, Neubig, and Liu, 2021) and GPTScore (Fu et al., 2023) that use BART or GPT models to evaluate the generated text.

3 CommonGen Task

We are going to compare the performance of different NLG architectures and study the suitability of some state-of-the-art metrics in the *CommonGen* task[1] (Lin et al., 2020). The objective is to create a coherent sentence describing an everyday situation using a set of given concepts. This task assesses both the capacity to apply commonsense knowledge and the ability to generalise when working with unfamiliar combinations of concepts.

CommonGen released a dataset to address this task, split into three subsets: training, validation and testing. For every set of concepts, several target sentences are provided. Table 1 shows the total number of concept sets and target sentences for each subset.

Corpus	*Train*	*Dev*	*Test*
Concept sets	32 651	993	1 497
Target sentences	67 389	4 018	N/A

Table 1: Commongen Dataset distribution.

To train our models we used the train subset. As reference sentences in the test subset are not available, to test the models' performance we used the evaluation subset.

4 Experimental Setup

To analyse the efficacy of diverse NLG automatic evaluation metrics, we have formulated a meticulously controlled scenario where the generated text can be evaluated. This scenario centres around addressing the research questions outlined in Section 1.

4.1 NLG Models

We wanted to compare the performance of classical architectures against the recent architectures; thus, we selected three different models (one modular architecture and two global approaches) to conduct the experimentation. We omitted to explore planning perspective architectures, which share similarities with modular architectures. The key distinction lies in the varying number of steps addressed during generation.

The first model we tested is SimpleNLG[2] (Gatt and Reiter, 2009), which handles the final step of language generation in the traditional modular NLG architecture, surface realisation. It is one of the most popular traditional NLG systems, and due to its popularity, it has been adapted to other languages, such as German (Braun et al., 2019), Mandarin (Chen, van Deemter, and Lin, 2018), Spanish (Ramos-Soto, Janeiro-Gallardo, and Bugarín, 2017) or Galician (Cascallar-Fuentes, Ramos-Soto, and Bugarín Diz, 2018). Since we already knew what information to incorporate into the final text (specifically, the concepts among the concept set), we followed an overgeneration and ranking strategy to obtain the syntactic representation. We first systematically generated all the possible sentence combinations of the three concepts. Then, these combinations were passed to SimpleNLG to perform the generation. Finally, we evaluated and ranked the resulting sentences according to Rouge_L (Lin, 2004) against the target sentences on the evaluation set to obtain the most suitable candidate sentence with a higher score.

The second model we used was trained using a fine-tuning strategy from a T5 model (Raffel et al., 2020). T5 is a pre-trained neural model that follows an encoder-decoder

[1]This task can be accessed on https://inklab.usc.edu/CommonGen/index.html.

[2]This API is available at https://github.com/simplenlg/simplenlg.

architecture. This model can perform well on various NLG tasks, by adding a prefix to the input. Furthermore, this model has shown great results on concept-to-text tasks (Roos, 2022), similar to the *CommonGen* task. We fine-tuned a T5-small pre-trained model, adding the tag *"CommonGen:"* before every input concept-set. Table 2 shows the hyperparameter configuration we used to perform the training.

Parameters	Values
Number train epochs	4
Batch size	16
Dropout	0.1
Learning rate	$1e-4$
Weight decay	0

Table 2: T5 fine-tuning hyperparameters.

The last model we tested was RMT (Zhang et al., 2023). We reproduced the model available at https://github.com/littlehacker26/Residual_Memory_Transformer because it showed promising results in addressing the CommonGen task and was publicly available. This model passes the probabilities of a GPT-2 decoder (Radford et al., 2019) to an encoder-decoder architecture with the novelty that this architecture incorporates three attention layers in the decoder, obtaining the last hidden states from the GPT-2 decoder, and the RMT encoder.

4.2 Metrics

We used different metrics to measure the performance of the aforementioned models. On the one hand, as CommonGen is a shared task with a leaderboard[3], we employed the same metrics as the leaderboard. This allowed us to somehow compare our results with those obtained by other participants. However, it is important to note that competitors' results were evaluated using the test subset (not publicly available), whereas the results from our experimentation are based on the evaluation subset, so a direct comparison is not possible. Those metrics are:

- *BLEU* (Papineni et al., 2002): This metric is commonly employed for machine translation tasks. It quantifies the word overlap between a candidate sentence and a target sentence, resulting in a score between 0 and 1. A higher value indicates greater similarity between the candidate sentence and the target sentence. In the *CommonGen* task, generated sentences were evaluated using BLEU_4. That means that it is calculated the precision of the word overlapping performs at a 4-gram level. This metric is configurable, making it possible to calculate the precision at 3-gram, 2-gram, and 1-gram level.
- *CIDEr* (Vedantam, Lawrence Zitnick, and Parikh, 2015): This metric was proposed to address the evaluation of image captioning[4]. Specifically, it evaluates the agreement between a candidate and a target sentence. To do so, first, perform the stemming of all words for both candidate and target sentence. Then, it measures the co-existence frequency of n-grams for both sentences, computing the weight for each n-gram using the Term Frequency Inverse Document Frequency (TF-IDF). Finally, it combines the scores of the different n-grams.
- *SPICE* (Anderson et al., 2016): This metric is commonly used on image captioning tasks as well. It measures the similarity between two sentences using the scene graph tuples parsed from the candidate sentence and the target sentences. Spice is computed based on the F1-Score between the tuples of candidate sentences and target sentences.

Moreover, to have a wider vision of the obtained results, we employed several additional metrics. These are:

- *Cosine Similarity*: This measurement quantifies the similarity between two non-zero vectors in an inner product space (Han, Kamber, and Pei, 2012). To do so, Candidate and target sentences are converted into a vector. Then, the cosine of the angle between these vectors is applied, resulting in a value ranging from -1 (indicating opposite directions)

[3]Leaderboard is available at https://inklab.usc.edu/CommonGen/leaderboard.html.

[4]Target sentences from the CommonGen dataset are extracted from image captioning datasets, and therefore, they consider that is more convenient to use metrics focused on image captioning tasks as they usually assume system generations and human references use similar concepts, and thus focus on evaluate the associations between mentioned concepts.

to 1 (representing highly similar directions). This metric can lead to more accurate results when comparing texts of varying sizes because it considers the angle between vectors in a dimensional space. (Guo, 2022) validates the suitability of cosine similarity to address the textual similarity evaluation.

- *ROUGE* (Lin, 2004): Originally, this metric was proposed to address text summarisation tasks. Despite its nature, ROUGE is one of the most employed metrics in NLG to measure the n-gram lexical overlap between the candidate and target sentences (Zhu and Bhat, 2020). ROUGE calculates the recall score of the candidate sentence corresponding to the target sentence. Specifically, Rouge-L identifies the longest co-occurring in sentence n-grams.

- *Flesch Reading Ease* (Kincaid et al., 1975): The evaluation criterion consists of determining how easy something is to read. Specifically, shorter words and shorter sentences will be easier to read. In the CommonGen task generated sentences should be dramatically simple and describe an everyday scenario. Thus, generating sentences should be easy to read. This metric produces a value between 0 and 121, being easier to read the higher the value is. In our experimentation, the results obtained within this metric are normalised in a range from 0 to 1.

- *BERTScore* (Zhang et al., 2020): This metric evaluates the quality of a text by performing the sum of cosine similarities between candidate and target sentences. Specifically, BERTScore obtains the contextual embeddings of those candidates and target sentences from BERT and calculates the cosine similarity across their tokens. This metric shows to correlate better with human judgements on sentence-level evaluation.

5 *Results and Discussion*

Within this section, we will expose the results obtained in our experimentation, analysing them in detail.

5.1 CommonGen Leaderboard Results

Firstly, to provide a brief background of the overall results obtained in the CommonGen task, we selected some of the models participating in it. Specifically, the following models are selected to report their performance according to their official results[5]: The best-performing model of the competition, DKMR2 (He et al., 2022). The worst-performing model of the competition, a fine-tuning of a T5-base model (Raffel et al., 2020). A model with intermediate results, a fine-tuning of a T5-large (Raffel et al., 2020). Table 3 shows the results obtained by these models on the test set.

Model	SPICE	CIDEr	BLEU_4
DKMR2	0.5243	0.3764	0.4649
T5-Large	0.2885	0.1512	0.3196
T5-Base	0.1987	0.0940	0.1854

Table 3: CommonGen leaderboard results.

Although DKMR2 is the top-performing model, it achieves discrete results across these metrics. Its highest score, 0.5243, is in the SPICE metric. However, DKMR2 significantly outperforms both T5-Large and T5-Base, nearly doubling their performance in all three metrics (SPICE, CIDEr, and BLEU).

Another aspect that needs to be taken into account is that these metrics seem to be aligned with the expected results achieved by these models. T5-Base obtains lower results compared to its larger version, T5-Large. This means that metrics are performing as expected for the entire test set.

5.2 NLG Architectures Results

This subsection aims to answer the research question: *Which type of NLG architectures perform best for commonsense generation?* To test the metrics evaluation effectiveness, we trained and reproduced the models explained in Section 4.1. With the models outlined in Section 4.1 we generated the sentences for the whole evaluation set, as that set contained a collection of target sentences to compare with. Table 4 shows the results obtained for the different metrics studied in this research work.

[5]https://inklab.usc.edu/CommonGen/.

Model	SPICE	CIDEr	BLEU_1	ROUGE-L	Readability	Cosine	BERTScore
SimpleNLG	0.183	0.059	0.129	0.310	0.139	0.374	0.900
T5-Small	0.256	0.109	0.600	0.444	0.215	0.287	0.914
RMT	0.215	0.024	0.371	0.345	0.343	0.302	0.889

Table 4: Results obtained by the experimented models.

T5-Small achieves the best results of the three tested models with the metrics used in the competition, SPICE, CIDEr, and BLEU. They are based on a word, or character overlapping among the candidate and the reference texts. In the same line, BERTScore and ROUGE-L also measure the embedding and word overlapping of both texts respectively. Consequently, they produce the same score, being T5-Small the best-performing model. Nonetheless, when comparing the cosine similarity of the target text against reference texts, SimpleNLG achieves a better score, as this metric does not penalise the length difference between texts. That means a shorter sentence could obtain a good score in this metric. Furthermore, the best score on the readability metric is obtained by the RMT model. That could indicate that sentences generated by RMT use a simpler vocabulary.

Otherwise, although the testing set was different from the test set evaluated in the competition, the results achieved by our experimented models are far from the best-performing model in the competition, DKMR2. However, when comparing with the other two selected models (T5-Base and T5-Large), the results are not as far. T5-Small scores 0.256 and 0.109 in the SPICE and CIDEr metrics respectively, while T5-Large obtains 0.288 and 0.151 in these metrics. That indicates, that the results are aligned with what is expected, being a little superior to the results obtained by a larger model. Differently, the results obtained by SimpleNLG —a traditional surface realisation engine— (0.183 and 0.059 in SPICE and CIDEr respectively), are relatively close to the results obtained by T5-Base (0.1987 and 0.0940), a model that follows a Transformer architecture, being SimpleNLG easier to use and less costly to train than T5.

In general, neural models (T5-Small and RMT) have performed better than a traditional architecture (SimpleNLG). Among both neural models, the best results are achieved by the T5-Small model, which according to our experiments and results, may be the most appropriate model to use for commonsense generation. Nevertheless, the improvement just by using these metrics is not as significant as the computing expense these models consume against a traditional architecture. Notwithstanding, analysing the performance of these models just by evaluating the generated sentences globally with automatic metrics might be not enough to be able to compare that improvement. Therefore, we will show a more detailed analysis.

5.3 Qualitative Analysis

Within this section, the research question we want to answer is: "*How good are the generated sentences of the best-performing system from a qualitative perspective?*". For this, we conducted a manual analysis of the entire set of sentences generated by SimpleNLG, T5-Small, and RMT models. Sentences generated with SimpleNLG tend to be a short combination of three concepts. In contrast, sentences generated by T5-Small and RMT are syntactically correct sentences, but sometimes they are semantically not accurate. However, they often repeat information within the same sentence.

To analyse those results deeply, we selected several examples of the generated sentences for different concept sets to make a detailed qualitative and human analysis. This selection was made based on the model with better results from the previous quantitative analysis, conducted in Section 5.2, i.e., T5-Small. To illustrate its qualitative performance, we selected a semantically correct sentence, a sentence that contained a fragment that was nonsensical, and a sentence which has no semantically correct fragments. Then, we selected the corresponding sentences of the same concept set in the other two models, SimpleNLG and RMT. Figure 2 shows the selected concept set and the sentence each model has generated for that collection of words.

As can be seen, the generated sentences by SimpleNLG are short and formed just by combining the words of the concept set.

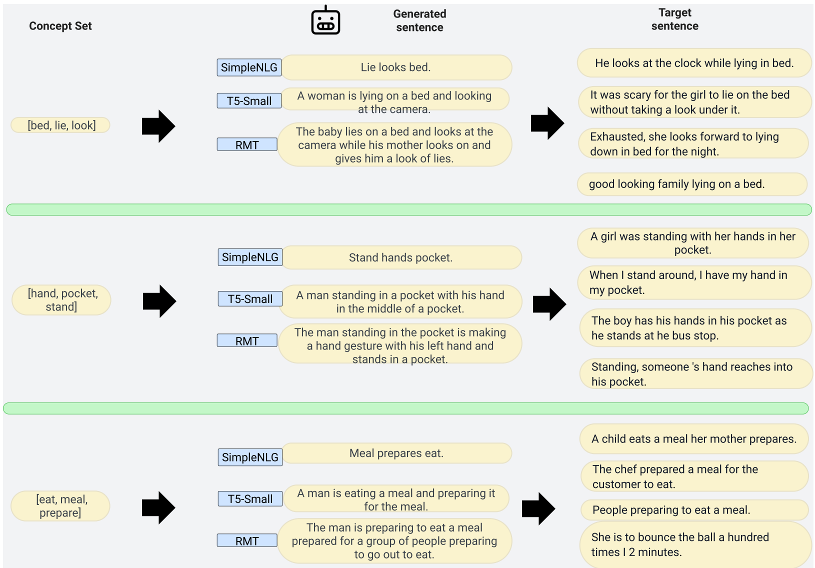


Figure 2: Results obtained for selected tuples.

These sentences are neither semantically nor syntactically correct, and a major part of the generated sentences by this model are nonsensical. In contrast, sentences generated by T5-Small are syntactically well formed and in most cases, they include commonsense knowledge. However, there are fragments in those sentences that are semantically incorrect, as the example "*A man is eating a meal and preparing it for the meal.*" that the first part of the sentence, (a man is eating a meal) do not correlate well with the second part (preparing it for the meal). Finally, sentences generated by RMT are also generally syntactically correct. Even so, they tend to repeat already said information in the generated sentence, e.g. the sentence *"The man standing in the pocket is making a hand gesture with his left hand and stands in a pocket."* repeats the information that the man is standing in a pocket. Moreover, some of these sentences are also nonsensical, such as the fragment "*gives him a look of lies*".

Table 5 shows the results obtained in the NLG evaluation metrics for the aforementioned generated sentences. For the first concept set, *"[bed, lie, look]"*, the model that better results achieves in most metrics is T5, as the sentence generated by it is better formed in terms of semantic and syntactic quality. Nevertheless, RMT and SimpleNLG sentences obtain better scores than T5 on the SPICE and BERTScore metrics, despite they are nonsensical. This suggests that these metrics may not be fully effective for evaluating this sentence.

Results obtained for the second concept set, *"[hand, pocket, stand]"* reveal a divergence among the generated sentences. While SPICE, BLEU_1, and ROUGE_L favour the sentence produced by T5, Cosine Similarity, readability and BERTScore indicate that the RMT-generated sentence is better. Upon closer analysis, both sentences exhibit correct syntax, but the sentence generated by RMT redundantly reiterates information already present in the sentence. So, this sentence is considered semantically worse than the sentence generated by T5.

Finally, all metrics suggest that the sentence generated by RMT achieves the best performance on the third concept set, *"[eat, prepare, meal]"*. Nonetheless, all the sentences are semantically inaccurate. SimpleNLG's generated sentence comment that a meal prepares to eat when that is not possible. In this respect, it is worth noting that sentences generated by SimpleNLG were configured to be in present tense only for simplicity reasons, as it would be very difficult to know a priori which verb tense would be more appropriate for each sentence for each given set of concepts. However, if this information was known, it would be possible to configure SimpleNLG to adjust the sentence to a specific tense, so in this example, the same sentence put in passive voice "Meal is prepared to eat" would be correct. T5's

Concept set	Model	SPICE	BLEU_1	Rouge-L	Readability	Cosine	BERTScore
[bed, lie, look]	**SimpleNLG**	0.133	0.264	0.306	0.167	0.293	0.896
	T5-Small	0.105	0.500	0.405	0.232	0.348	0.893
	RMT	0.154	0.391	0.311	0.159	0.254	0.895
[hand, pocket, stand]	**SimpleNLG**	0.095	0.189	0.280	0.226	0.323	0.876
	T5-Small	0.133	0.471	0.389	0.265	0.322	0.905
	RMT	0.129	0.450	0.310	0.315	0.368	0.908
[eat, meal, prepare]	**SimpleNLG**	0.143	0.368	0.336	0.226	0.374	0.909
	T5-Small	0.105	0.429	0.344	0.240	0.253	0.921
	RMT	0.211	0.500	0.482	0.366	0.368	0.928

Table 5: Results obtained for the selected concept sets.

sentence contains the actions of eating and preparing a meal at the same time when that action can not be taken simultaneously. Finally, RMT's sentence mentions that a man is eating a meal prepared for another group of people. This sentence is not completely semantically wrong but is difficult to understand. It also repeats some information, the action of preparing to eat, so this issue affects the quantitative results obtained by this sentence that may increase. One aspect to remark on within this concept set is that a confusing sentence such as the one generated by RMT, obtains the best score on the readability score, suggesting that this metric may be not accurate in this generation context.

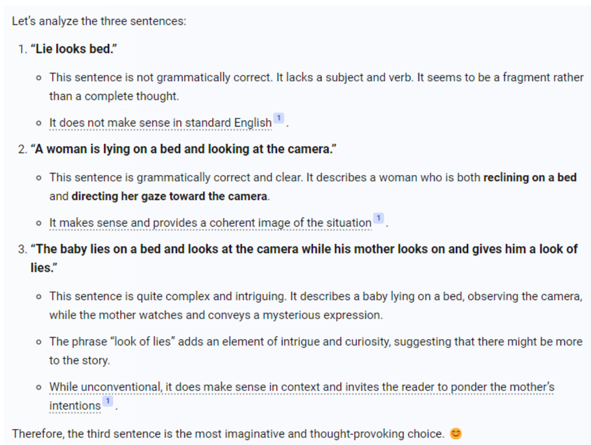
Generally speaking, since almost all the evaluation metrics highly rely on the target sentences with which the generated sentences are compared, the best option to be confident in the quantitative results achieved by a model is to compare different metrics and check the correlation among them, not forgetting the qualitative evaluation of the generated sentences as well. As shown in our empirical analysis, T5-Small, the model with better quantitative results, generated syntactically correct sentences which sometimes contained semantically wrong fragments.

5.4 ChatGPT as a Text Evaluator

To debate the research question "*To what extent can LLMs-based tools, such as ChatGPT, be a good alternative or complement to automatic evaluation metrics?*" we conducted a preliminary experimentation using the same concept sets as in Section 5.3.

We prompted GPT-4 model (OpenAI, 2023) via Bing interface[6] the following text: *"Given the following three sentences, which one does make sense?"*. After that, the generated sentences are written. GPT-4 gives as output a detailed explanation of why each of the sentences makes sense or not, and argues which sentence is the most appropriate.

Figure 3 shows the generated output for the sentences generated with the concept set *[bed, lie, look]*. GPT-4 argues that the sentence generated with SimpleNLG is not grammatically correct, correlating with human judgement. For the sentences generated with T5-Small and RMT respectively, it says that they are grammatically correct. However, analysing the RMT sentence, the fragment *"gives him a look of lies"* is not semantically accurate. Despite this, GPT-4 concludes that the third sentence including that fragment is the best choice.


Let's analyze the three sentences:

1. **"Lie looks bed."**
   - This sentence is not grammatically correct. It lacks a subject and verb. It seems to be a fragment rather than a complete thought.
   - It does not make sense in standard English 1.
2. **"A woman is lying on a bed and looking at the camera."**
   - This sentence is grammatically correct and clear. It describes a woman who is both **reclining on a bed** and **directing her gaze toward the camera**.
   - It makes sense and provides a coherent image of the situation 1.
3. **"The baby lies on a bed and looks at the camera while his mother looks on and gives him a look of lies."**
   - This sentence is quite complex and intriguing. It describes a baby lying on a bed, observing the camera, while the mother watches and conveys a mysterious expression.
   - The phrase "look of lies" adds an element of intrigue and curiosity, suggesting that there might be more to the story.
   - While unconventional, it does make sense in context and invites the reader to ponder the mother's intentions 1.

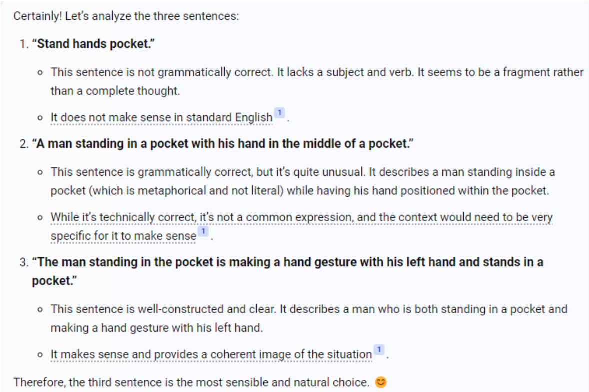
Therefore, the third sentence is the most imaginative and thought-provoking choice. 😊

Figure 3: Generated sentence for the concept set [bed, lie, look].

The output generated for the second concept set (*[hand, pocket, stand]*) can be seen in Figure 4. In this output, GPT-4 contradicts itself on the action of *"a man standing in a pocket"*. While the second sentence says that this action is not correct (it is metaphorical), it argues that the third sentence is well-constructed and clear. More-

[6] GPT-4 can be accessed at https://www.bing.com/chat

over, GPT-4 concludes that the third sentence is the most natural choice, when it repeats the metaphorical action of "standing in a pocket" twice. At last, the concept set

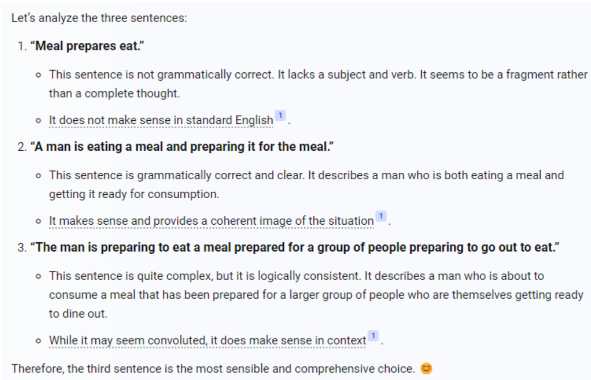

Certainly! Let's analyze the three sentences:

1. **"Stand hands pocket."**
   - This sentence is not grammatically correct. It lacks a subject and verb. It seems to be a fragment rather than a complete thought.
   - It does not make sense in standard English [1].
2. **"A man standing in a pocket with his hand in the middle of a pocket."**
   - This sentence is grammatically correct, but it's quite unusual. It describes a man standing inside a pocket (which is metaphorical and not literal) while having his hand positioned within the pocket.
   - While it's technically correct, it's not a common expression, and the context would need to be very specific for it to make sense [1].
3. **"The man standing in the pocket is making a hand gesture with his left hand and stands in a pocket."**
   - This sentence is well-constructed and clear. It describes a man who is both standing in a pocket and making a hand gesture with his left hand.
   - It makes sense and provides a coherent image of the situation [1].

Therefore, the third sentence is the most sensible and natural choice. 😊

Figure 4: Generated sentence for the concept set [hand, pocket, stand].

[eat, meal, prepare], although being accurate in the conclusion that the last sentence is the most appropriate, fails at evaluating the second sentence. GPT-4 says that this sentence makes sense. But eating a meal while preparing it, is not physically possible, so that reasoning is not valid. As we have seen, GPT-4


Let's analyze the three sentences:

1. **"Meal prepares eat."**
   - This sentence is not grammatically correct. It lacks a subject and verb. It seems to be a fragment rather than a complete thought.
   - It does not make sense in standard English [1].
2. **"A man is eating a meal and preparing it for the meal."**
   - This sentence is grammatically correct and clear. It describes a man who is both eating a meal and getting it ready for consumption.
   - It makes sense and provides a coherent image of the situation [1].
3. **"The man is preparing to eat a meal prepared for a group of people preparing to go out to eat."**
   - This sentence is quite complex, but it is logically consistent. It describes a man who is about to consume a meal that has been prepared for a larger group of people who are themselves getting ready to dine out.
   - While it may seem convoluted, it does make sense in context [1].

Therefore, the third sentence is the most sensible and comprehensive choice. 😊

Figure 5: Generated sentence for the concept set [eat, meal, prepare].

explains why a sentence is valid or not. Nevertheless, that explanation is not always correlated with human judgements, and neither is accurate with its explanations. Therefore there is still room for improvement in the way LLMs such as GPT-4 evaluate the text appropriateness.

6 Conclusions and Future Work

This paper presented an empirical analysis of several NLG evaluation metrics and models for the commonsense generation task. The experiments conducted and the discussion of the results led us to conclude that current evaluation metrics highly depend on the set of target sentences a text is compared with.

As we have seen, the alignment between automated metrics and human evaluation is not always accurate. When comparing the performance of different models, the results obtained by SimpleNLG are similar to the results achieved by the other two models. However, it is important to note that, when analysing these sentences manually, the sentences generated by SimpleNLG are just a combination of concepts, often resulting in nonsensical and not syntactically correct sentences, while the sentences generated by the other two models are at least syntactically correct. This difference is not reflected in the results obtained in these metrics. This may indicate that using words that are contained in the target sentences in a random order, can produce similar results to syntactically correct sentences using these metrics.

Another important point to consider when evaluating using the standard NLG evaluation metrics is repetition. In natural language, repeating the same information can be redundant, making that sentence of inferior quality to one that avoids the repetition. Nevertheless, in the context of these metrics, the repetition not only fails to penalise the results obtained but raises the results obtained. Furthermore, we have seen that some sentences that are not semantically correct obtain better results than sentences that are semantically accurate, as most used metrics do not evaluate the semantic information.

Finally, evaluating a model with several metrics is more accurate than doing so with only one metric, as it gives a wider vision of how that model is performing.

One future line of work is to expand the preliminary analysis of LLMs as text evaluators, exploring different available LLMs, as they can capture and learn human patterns from a semantic perspective. Therefore, these models can indeed learn such patterns, they may serve as valuable tools for evaluating sentence correctness. Additionally, it would be interesting to explore how to enhance the performance of our top-performing model, T5-Small, also by incorporating commonsense knowledge into the model.

Acknowledgements

The research work conducted is part of the R&D projects "CORTEX: Conscious Text Generation" (PID2021-123956OB-I00), funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe"; "CLEAR.TEXT:Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities" (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and "European Union NextGenerationEU/PRTR"; and the project "NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/21)" funded by the Generalitat Valenciana. Moreover, it has been also partially funded by the European Commission ICT COST Action "Multi-task, Multilingual, Multi-modal Language Generation" (CA18231).

References

Aghahadi, Z. and A. Talebpour. 2022. Avicenna: a challenge dataset for natural language generation toward commonsense syllogistic reasoning. *Journal of Applied Non-Classical Logics*, 32(1):55–71.

Anderson, P., B. Fernando, M. Johnson, and S. Gould. 2016. SPICE: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

Appelt, D. 1985. Planning english sentences. cambridge university press.

Banerjee, S. and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Bateman, J. A. 1997. Enabling technology for multilingual natural language generation: the KPML development environment. *Natural Language Engineering*, 3(1):15–55.

Bhargava, P. and V. Ng. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12317–12325.

Braun, D., K. Klimt, D. Schneider, and F. Matthes. 2019. SimpleNLG-DE: Adapting SimpleNLG 4 to German. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 415–420, Tokyo, Japan, October–November. Association for Computational Linguistics.

Carlsson, F., J. Öhman, F. Liu, S. Verlinden, J. Nivre, and M. Sahlgren. 2022. Fine-grained controllable text generation using non-residual prompting. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6837–6857.

Cascallar-Fuentes, A., A. Ramos-Soto, and A. Bugarín Diz. 2018. Adapting SimpleNLG to Galician language. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 67–72, Tilburg University, The Netherlands, November. Association for Computational Linguistics.

Chen, G., K. van Deemter, and C. Lin. 2018. SimpleNLG-ZH: a linguistic realisation engine for Mandarin. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 57–66, Tilburg University, The Netherlands, November. Association for Computational Linguistics.

Dong, C., Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang. 2023. A survey of natural language generation. *ACM Computing Surveys*, 55:1–38, 8.

Fu, J., S.-K. Ng, Z. Jiang, and P. Liu. 2023. GPTScore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Gatt, A. and E. Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Gatt, A. and E. Reiter. 2009. SimpleNLG: A realisation engine for practical applications. In *Proceedings of the 12th European* *workshop on natural language generation (ENLG 2009)*, pages 90–93.

Guo, K. 2022. Testing and validating the cosine similarity measure for textual analysis. *Available at SSRN 4258463*.

Han, J., M. Kamber, and J. Pei. 2012. 2 - getting to know your data. In *Data Mining (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Boston, third edition edition, pages 39–82.

He, X., Y. Gong, A.-L. Jin, W. Qi, H. Zhang, J. Jiao, B. Zhou, B. Cheng, S. Yiu, and N. Duan. 2022. Metric-guided distillation: Distilling knowledge from the metric to ranker and retriever for generative commonsense reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 839–852, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Hovy, E. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Khapra, M. M. and A. B. Sai. 2021. A tutorial on evaluation metrics used in natural language generation. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Tutorials*, pages 15–19.

Kincaid, J. P., R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Millington (Tenn.)*.

Koller, A. and M. Stone. 2007. Sentence generation as a planning problem. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 336–343, Prague, Czech Republic, June. Association for Computational Linguistics.

Kusner, M., Y. Sun, N. Kolkin, and K. Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Lemon, O. 2011. Learning what to say and how to say it: Joint optimisation of spoken dialogue management and natural language generation. *Computer Speech & Language*, 25(2):210–221.

Levelt, W. 1989. Speaking: From intention to articulation MIT press. *Cambridge, MA*.

Lin, B. Y., W. Zhou, M. Shen, P. Zhou, C. Bhagavatula, Y. Choi, and X. Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online, November. Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Lo, C.-k., A. K. Tumuluru, and D. Wu. 2012. Fully automatic semantic MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 243–252.

Mann, W. C. and J. A. Moore. 1981. Computer generation of multiparagraph english text. *American Journal of Computational Linguistics*, 7(1):17–29.

McDonald, D. D. 2010. Natural language generation. *Handbook of natural language processing*, 2:121–144.

Mirza, M., B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, I. J. Goodfellow, and J. Pouget-Abadie. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680.

Nakatsu, C. and M. White. 2010. Generating with discourse combinatory categorial grammar. *Linguistic Issues in Language Technology*, 4.

Nirenburg, S., V. R. Lesser, and E. Nyberg. 1989. Controlling a language generation planner. In *IJCAI*, pages 1524–1530.

OpenAI. 2023. GPT-4 technical report.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text Transformer. *J. Mach. Learn. Res.*, 21(1), jan.

Ramos-Soto, A., J. Janeiro-Gallardo, and A. Bugarín. 2017. Adapting SimpleNLG to spanish. pages 144–148. Association for Computational Linguistics.

Reiter, E. 1994. Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on Natural Language Generation*.

Rieser, V. and O. Lemon. 2009. Natural language generation as planning under uncertainty for spoken dialogue systems. *Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation*, pages 105–120.

Roos, Q. 2022. Fine-tuning pre-trained language models for CEFR-level and keyword conditioned text generation: A comparison between google's t5 and openai's gpt-2.

Sai, A. B., A. K. Mohankumar, and M. M. Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.*, 55(2), jan.

Scarselli, F., M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

Stanchev, P., W. Wang, and H. Ney. 2019. EED: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 514–520.

Sukhbaatar, S., J. Weston, R. Fergus, et al. 2015. End-to-end memory networks. *Advances in neural information processing systems*, 28.

Sutskever, I., O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Tang, T., H. Lu, Y. E. Jiang, H. Huang, D. Zhang, W. X. Zhao, and F. Wei. 2023. Not all metrics are guilty: Improving NLG evaluation with LLM paraphrasing. *arXiv preprint arXiv:2305.15067*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vedantam, R., C. Lawrence Zitnick, and D. Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Wang, H., Y. Liu, C. Zhu, L. Shou, M. Gong, Y. Xu, and M. Zeng. 2021. Retrieval enhanced model for commonsense generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3056–3062, Online, August. Association for Computational Linguistics.

Wang, J., Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, and J. Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore, December. Association for Computational Linguistics.

Yu, W., C. Zhu, L. Qin, Z. Zhang, T. Zhao, and M. Jiang. 2022. Diversifying content generation for commonsense reasoning with mixture of knowledge graph experts. In *NAACL 2022 Workshop on Deep*

Learning on Graphs for Natural Language Processing.

Yuan, W., G. Neubig, and P. Liu. 2021. BARTScore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Zhang, H., S. Si, H. Wu, and D. Song. 2023. Controllable text generation with residual memory transformer. *arXiv preprint arXiv:2309.16231.*

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations.*

Zhang, Y. and X. Wan. 2024. SituatedGen: Incorporating geographical and temporal contexts into generative commonsense reasoning. *Advances in Neural Information Processing Systems*, 36.

Zhu, W. and S. Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 94–108, Online, November. Association for Computational Linguistics.

Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 137-148 recibido 15-12-2023 revisado 30-01-2024 aceptado 04-02-2024

OntoLM: Integrating Knowledge Bases and Language Models for classification in the medical domain

OntoLM: Integrando bases de conocimiento y modelos de lenguaje para clasificación en dominio médico

Fabio Yáñez-Romero[1], Andres Montoyo[2], Rafael Muñoz[2], Yoan Gutiérrez[2], Armando Suárez[2]

[1]University Institute for Computer Research, University of Alicante.
[2]Department of Computing and Information Systems, University of Alicante.

fabio.yanez@ua.es, montoyo@dlsi.ua.es, rafael@dlsi.ua.es, ygutierrez@dlsi.ua.es, armando@dlsi.ua.es

Abstract: Large language models have shown impressive performance in Natural Language Processing tasks, but their black box characteristics render the explainability of the model's decision difficult to achieve and the integration of semantic knowledge. There has been a growing interest in combining external knowledge sources with language models to address these drawbacks. This paper, *OntoLM*, proposes a novel architecture combining an ontology with a pre-trained language model to classify biomedical entities in text. This approach involves constructing and processing graphs from ontologies and then using a graph neural network to contextualize each entity. Next, the language model and the graph neural network output are combined into a final classifier. Results show that *OntoLM* improves the classification of entities in medical texts using a set of categories obtained from the Unified Medical Language System. We can create more traceable natural language processing architectures using ontology graphs and graph neural networks.
Keywords: External Knowledge, Ontologies, Large Language Models, Graph Neural Networks.

Resumen: Los grandes modelos de lenguaje han mostrado un rendimiento impresionante en tareas de Procesamiento del Lenguaje Natural, pero su condición de caja negra hace difícil explicar las decisiones del modelo e integrar conocimiento semántico. Existe un interés creciente en combinar fuentes de conocimiento externas con LLMs para solventar estos inconvenientes. En este artículo, proponemos *OntoLM*, una arquitectura novedosa que combina una ontología con un modelo de lenguaje pre-entrenado para clasificar entidades biomédicas en texto. El enfoque propuesto consiste en construir y procesar grafos provenientes de una ontología utilizando una red neuronal de grafos para contextualizar cada entidad. A continuación, combinamos los resutlados del modelo de lenguaje y la red neuronal de grafos en un clasificador final. Los resultados muestran que *OntoLM* mejora la clasificación de entidades en textos médicos utilizando un conjunto de categorías obtenidas de Unified Medical Language System. Utilizando grafos de ontologías y redes neuronales de grafos podemos crear arquitecturas de procesamiento de lenguaje natural más rastreables.
Palabras clave: Conocimiento Externo, Ontologías, Grandes Modelos de lenguaje, Redes Neuronales de Grafos.

ISSN 1135-5948 DOI 10.26342/2024-72-10

1 Introduction

This work is centred on the premise that using structured external knowledge can help during the fine-tuning process of large language models, and it also makes the architecture more traceable and explainable as it provides semantic knowledge during the process. To validate the premise, a multilabel classification task is chosen. In this task structured knowledge is used with language models forming an even larger architecture which combines the language model with a graph neural network (GNN) in a final classifier.

This work aims to insert structured external knowledge into the decision-making of a model based on pretrained language models, improving the results obtained in classification tasks and obtaining a final architecture ($OntoLM$) that will allow traceability through the GNN and the initial structures obtained from UMLS.

An ontology defines the possible relations between different types of entities and is used as a schema to decide how relational information should be stored in an ordered way. The rules defined in the ontology are expressed in the final knowledge base (KB) derived from this ontology. KBs store information about many domains in a structured way. Big KBs like UMLS or WordNet have proven their usefulness in many downstream tasks where factual information is needed, reducing the amount of wrong information returned (Chen et al., 2017). AlKhamissi et al. (2022) consider the following criteria as the most important characteristics for considering a language model as a KB:

- **Accessibility**: all the information of a KB can be queried directly.
- **Easy to edit**: every entity or relation can be modified with minor effort.
- **Consistency**: queries with the same meaning should give the same result.
- **Reasonableness**: how suitable is the application of reasoning techniques over their structure rather than deep learning models.
- **Explainability and interoperability**: explainable algorithms and techniques are more suitable; for example, knowledge base schema or path walking techniques.

By contrast, big deep learning models used in Natural Language Processing (NLP) store large amounts of information through their training with large amounts of text, as shown in the most recent cases with BERT (Devlin et al., 2019) or GPT-4 (OpenAI et al., 2024). Language models have proven to be very useful in numerous tasks carried out in language processing. Their different architectures allow them to cover both classification and text generation tasks. However, these models have a large amount of probabilistic knowledge which cannot be interpreted.

Using only language models can present problems because of the lack of internal reasoning in such a model, as well as biases (Bender et al., 2021) and toxic information (Gehman et al., 2020) contained in them. The information obtained in these models is not easy to update, so they tend to be easily outdated due to the high cost of re-training them. Also, these models have many inconsistencies, as shown by works that obtain different information using prompt engineering techniques (Elazar et al., 2021).

Finally, traceability, interpretability, and explainability are easier to achieve with a well-defined ontology that generates information based on certain rules or schema and their graph structure (Agarwal et al., 2023). Deep learning models that consider the entire structure of a graph in the training data often provide more traceable structures that can be understood intuitively (Zhou et al., 2020).

The paper is structured as follows: Section 2 discusses other works using similar approaches, trying to provide semantic knowledge with external knowledge or training data for language models. Section 3 describes the aim and characteristics of the corpus associated with the experiment. The next sections, 4 focuses on the whole architecture of the experiment with a specific focus on data processing and model training 5. Subsequently, the results are reported in section 6. The discussion of the results obtained is carried out in section 7, whereas conclusions and future work arising from the discussion are explained in section 8.

2 Related Work

The use of external knowledge and language models has been extensively researched to address the issues encountered in language models. Some works, such as (Kaur et al.,

2022) or (Sun et al., 2021), aim to pretrain the model by incorporating semantic knowledge or altering the existing architecture. Others train a language model from scratch using an innovative masking approach (Zhang et al., 2019), improving many benchmarks. In both cases, the computational cost is high, making it difficult to adopt similar experiment strategies.

Other approaches try to bring semantic knowledge into the language model without updating the language model parameters, either by using pre-processing (Sun et al., 2024) or post-processing (He, Zhang, and Roth, 2022) techniques. These approaches are usually less expensive, making them more accessible and versatile than previous examples.

The factual knowledge contributed to the language model can have an unstructured origin, as in the case of Peng et al. (2023), or it can come from structured knowledge bases, where the information is mainly organised in the form of triples (Huang et al., 2022).

The advantage of using a structured knowledge source as external knowledge is the elimination of ambiguities present in the text, as well as an ordered information structure that does not introduce more noise than necessary and the provision of semantic knowledge, such as synonymy, hyponymy, and hyperonymy or antonymy relations (Mrkšić et al., 2016) depending on the knowledge base used.

Previous work has attempted to provide structured knowledge by capturing the semantics of KBs and feeding this knowledge into deep learning models from modules specialised in this task (Piad-Morffis et al., 2019).

Other works have been carried out that attempt to benefit from the knowledge present in knowledge bases with GNNs because of the inherent relation of this architecture with their different nodes. Jiang et al. (2020) proposes to use knowledge from a graph to perform text classification, in their case they create the graph by performing Named Entity Recognition (NER) on short texts, augmenting the information obtained with a general knowledge base and initiating embeddings of each entity using Word2Vec (Mikolov et al., 2013). The graph is processed using Gated Graph Neural Networks (GGNNs) (Li et al., 2017), and they also process the whole text with a pre-trained language model (PTLM). Finally, they use an attention background on the GGNN and the PLM results to classify the text. There are other methods to create embeddings of the entities and relations of a knowledge graph. These methods can be considered contextualised embeddings from knowledge graphs (Yáñez Romero et al., 2023-09).

Another example is Feng et al. (2020), who use pre-trained language models and knowledge from different ontologies to answer questions with a fixed number of answers as context. In their work, they form different graphs with ontology entities from the entities detected in both question and possible answers. This information is passed through a GNN that considers the type of relation between each node and a node scoring system to filter the possible paths between questions and answers. This novel way of applying external knowledge to language models has a major problem: it is used specifically to respond to questions with a fixed number of answers.

In this proposal, an architecture similar to (Yasunaga et al., 2021) is used to classify entities detected in a given text. For this purpose, language models trained in the specific domain of the text and ontologies with knowledge of the same domain will be used. The architecture of the graphs used during training will be adapted to the proposed objective, and the GNN introduced by Feng et al. (2020) and the improvements introduced by Yasunaga et al. (2021) will be utilized.

3 Corpus

The corpus used to classify medical entities has been created by annotating medical terms found in abstracts of papers obtained from PubMed. Annotated texts focus on diseases, as these texts were collected to classify entities related to diseases and ailments. The annotations made contemplate 40 different categories obtained from the semantic types of UMLS. Specifically, this corpus has been annotated semi-automatically by performing NER on the abstracts using NER models obtained from sci-spacy (Neumann et al., 2019), namely 'en-core-sci-lg'. Then, each annotation was supervised, and labels that did not correspond to the context of the entity were removed. However, the corpus used is very unbalanced, as differences of 1 to 100 can be found between the categories with the lowest representation and those with the highest representation. This problem was mitigated by undersampling the dataset.

Finally, a corpus was obtained where each entity can be annotated with more than one category since, in its context, this entity can be considered within different UMLS categories, e.g. in Text 1 the entity *pharmacological treatments* can be classified as *healthcare activity* or *research activity*. Therefore, we are faced with a multi-label classification problem.

> *This work analyzed salivary Lf concentration under different handling conditions and donor-dependent factors, including age, inter-diurnal variations, physical activity, and* ***pharmacological treatments****.* (1)

However, in the corpus used, most entities are classified with only one label, with a few examples having two labels and almost none with more, as shown in Figure 1.

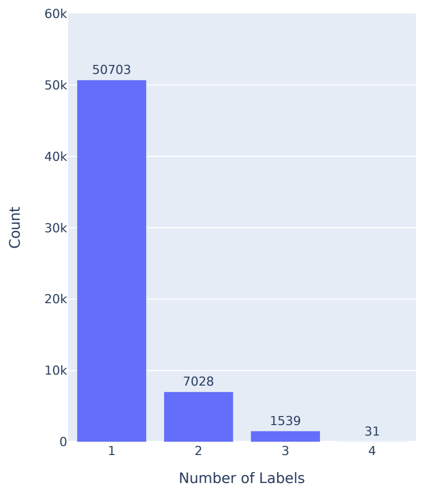


Figure 1: Clustering of training data based on the number of labels.

4 *OntoLM*

The proposal performs a supervised classification task on annotated medical entities obtained from biomedical texts. In brief, supervised learning is performed by augmenting a pre-trained language model with the knowledge from the UMLS medical ontology, carrying out the training of a GNN. The different sections of the architecture are listed as follows:

1. The starting point is the data obtained from the corpus with their corresponding labels.
2. The UMLS ontology database is then used to augment the annotated data with the possible entities detected in UMLS (4.1).
3. The graphs necessary for the model's training are created according to section 4.2.
4. The language models used to represent the entities and process the text are named in section 4.3.
5. Finally, a summary of the task covered in the experiment is given 4.5.

Figure 2 shows the complete proposed architecture, starting from the texts with annotated entities. In the pre-processing stage, each text document is represented in light yellow, the database and the entities extracted from it can be seen in orange, and the input tensors to the model are in grey. During the training step, depicted in light blue/salmon, the model's components are frozen/learning, respectively. Numerical values represent the order of the data flow.

The complete experiment considers all three components: GNN, LLM, and a multilayer perceptron (MLP).

4.1 Ontology structure

Understanding the structure of the initial ontology from which one starts is essential to forming a coherent graph for the proposed task. In this case, UMLS will be used as it contains much knowledge from the medical field (Bodenreider, 2004).

Considering UMLS entities and relations as a graph, the minimum structure of this database would be triples. Each triple consists of a head entity e_h and a tail entity e_t with its respective relations r so that a triple would be represented by the expression (e_h, r, e_t). It is possible to generate a knowledge graph from UMLS using specific data tables that indicate the relations between the different entities, i.e. triples.

These concepts and relations constitute the main data source of UMLS, known as its Metathesaurus. However, UMLS has other sources of knowledge, such as the Lexicon, which generates the different linguistic variants of a term, or the semantic network, which generates higher-level categories that encompass the concepts in the database. The

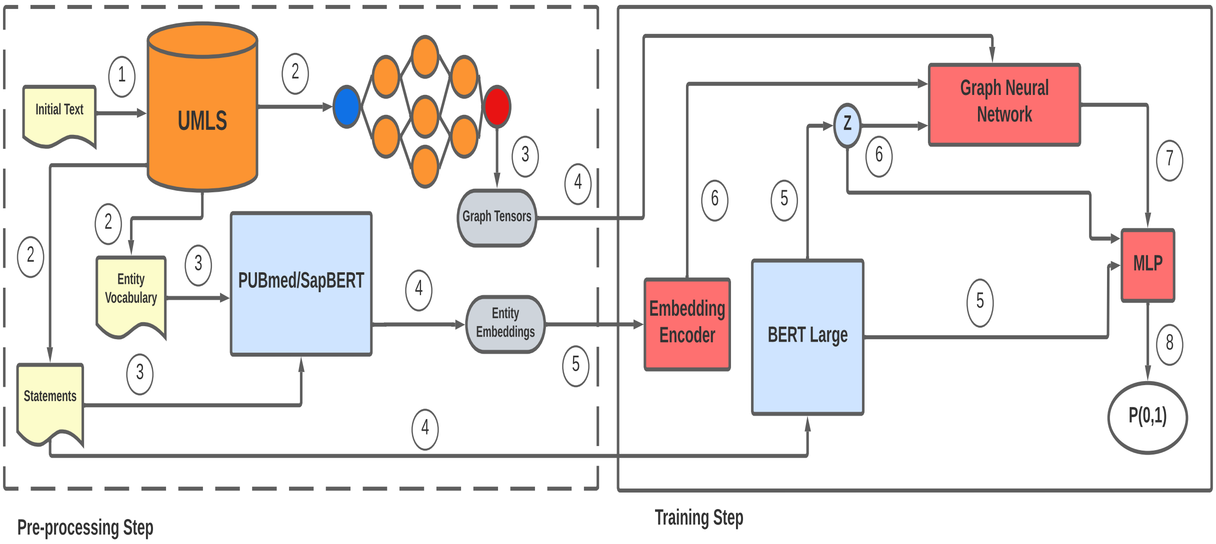


Figure 2: OntoLM architecture. The box above shows all the pre-processing of the data carried out before training. The bottom box shows the training stage.

UMLS semantic network classifies each concept based on Semantic Types (TUIs) (McCray, 1989), which can represent direct relations between the different concepts or classify these concepts in higher-level categories. Moreover, TUIs are organised hierarchically among themselves and can have a subset within another set.

The TUIs used to catalogue the different concepts can be used as categories in entity classification problems within the biomedical field. Considering all the classification TUIs, there are 127 different ones, forming a large number to be used in classification problems. Of the 127 initial categories obtained from UMLS, 34 were eliminated because they did not provide value for classifying disease-related entities, such as *Temporal concept* or *Geographical Area*. This leaves a total of 93 categories, a particularly high number for multilabel classification using language models. For this reason, the 93 categories have been reduced by grouping them by their hierarchical relations so that *Plant*, *Fungus* or *Animal* can be grouped under TUI *Organism*. The number of categories considered has been reduced to the 20 most representative ones to balance the final categories obtained. The 40 initial categories and the chosen 20 categories are shown in Table 1.

For each example to be classified, 20 graphs are generated with the detected entities of the text, which a GNN then processes. Also, 20 statements are generated and processed by the language model. A graph and a statement are generated for each possible category among all those considered. The construction method of each graph and statement is indicated in the following sections.

UMLS Id	Category Names	Nº Labels
T001	Organism	-
T005	Virus	-
T007	Bacterium	-
T018	Embryonic Structure	-
T023	**Body Part Organ Or Organ Component**	**801**
T025	**Cell**	**801**
T026	**Cell Component**	**801**
T028	**Gene Or Genome**	**801**
T032	Organism Attribute	-
T033	**Finding**	**801**
T037	Injury Or Poisoning	-
T038	**Biologic Function**	**801**
T043	**Cell Function**	**801**
T046	**Pathologic Function**	**801**
T047	**Disease or Syndrome**	**801**
T049	**Cell or Molecular Dysfunction**	**801**
T050	Experimental Model of Disease	-
T055	Individual Behavior	-
T058	**HealthCare Activity**	**801**
T062	**Research Activity**	**801**
T066	Machine Activity	-
T069	Environmental Effect of Humans	-
T070	Natural Phenomenon or Process	-
T073	Manufactured Object	-
T079	**Temporal Concept**	**801**
T085	Molecular Sequence	-
T091	Biomedical Occupation Or Discipline	-
T093	HealthCare Related Organization	-
T098	**Population Group**	**801**
T101	**Patient or Disabled Group**	**801**
T103	**Chemical**	**801**
T114	Nucleic Acid Nucleoside or Nucleotide	-
T116	**AminoAcid Peptide or Protein**	**801**
T121	**Pharmacologic Substance**	**801**
T123	**Biologically Active Substance**	**801**
T167	Substance	-
T184	Sign or Symptom	-
T190	Anatomical Abnormality	-
T201	**Clinical Attribute**	**801**
T204	Eukaryote	-
	Total Statements	**15321**

Table 1: The 40 initial categories considered in the classification task, and the 20 final categories used after undersampling the dataset.

4.2 Proposed Graph Structure

For the classification of words from a text, it is necessary to modify the network architecture proposed in (Feng et al., 2020) and (Yasunaga et al., 2021), since it is not about answering questions. Therefore the possible answers cannot be used as a context.

In this structure, there is an initial entity, which is the target entity to classify, and the rest of the entities detected in the text belonging to the biomedical field. The other entities detected will serve as context to classify the target word.

To introduce the context of the entity to be classified and the possible classification it refers to, entities that do not exist in UMLS are created representing the exact word found in the text, and new relations that will connect these entities with UMLS entities. For instance, the word 'results' in a medical text may refer to different entities within the knowledge base, such as 'Clinical results' or 'Experimental results'.

From the initial entity, which is the annotated entity, using matching with n-grams of three characters, the possible entities referred to by that word are obtained, each with their respective classifications. The initial entity is related to these entities from the ontology using a newly created relation *meaning of*. From the ontology entities, the rest of the entities directly connected to them that share one or more semantic types can be obtained using the UMLS database. This step can be done many times, increasing the size of the final graph. The final node of the network is each of the possible categories used in the architecture. This node will be directly connected to the rest of the entities detected in the target text related to the category (based on UMLS possible entities). The new relation used in this case will be *belongs to*. The considered relations can be expanded with direct relations to the possible TUIs of the word to be classified, further extending the graph and thus connecting to the initial entities. The intermediate triples obtained from the ontology present the different relations considered in the UMLS version.

Figure 3 illustrates the architecture of each graph; every node represents an entity, the blue nodes being the target entity in the text to be classified. The green nodes are the possible nodes obtained from the ontology. The orange nodes represent the remaining nodes obtained from the text that have a category that coincides with the possible categories of the green nodes (based on three characters n-gram matching on UMLS). The red nodes represent one of the 20 possible categories, which matches the orange nodes category. Finally, the white nodes represent those obtained from existing relations in UMLS with the rest of the previously mentioned nodes.

To avoid information loss and improve the results if many jumps are made, the contextual node Z is used (Yasunaga et al., 2021). This node connects the initial node (target word to be classified) with the final node (possible category).

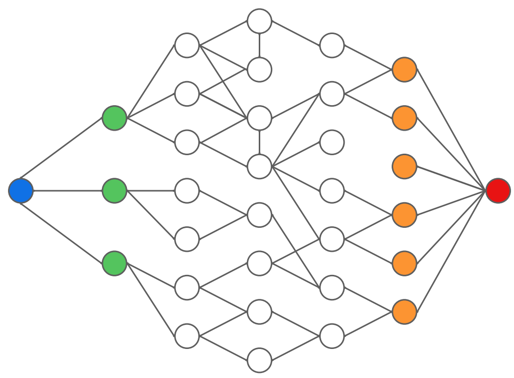


Figure 3: Graph structure proposed by each statement.

In cases where the target category is not related to the entities of the text, only the category is added as an isolated node. The other categories of the text are not added. Thus, a network is formed by the nodes obtained from the initial entity and the isolated node (connected only by the context node Z).

The proposed architecture is processed using the GNN introduced by Feng et al. (2020), then a pooling is performed on the GNN and fed into an MLP together with the language model data and the context node Z.

4.3 Language Models

To represent the nodes of each graph in a format compatible with the GNN, it is necessary to use embeddings containing the information of each entity. Each node is initialised using a specific language model for this task in this case.

Language models can be used to create entity embeddings, as they store a large amount of knowledge in their model weights. With this in mind, language models trained on a specific domain can represent entities and their relations from that domain for subsequent tasks. This is the case of other works such as (Wang et al., 2023) or (Wang et al., 2022), where entity embeddings are created using language models for entity linking and relation inference.

A BERT model trained with UMLS data is used to generate the embeddings of the medical entities. This model was trained to represent the different names that the same medical concept can have in a similar way, which is ideal for the present task (Liu et al., 2021). In addition, the SapBert model used is based on PubMedBERT (Gu et al., 2021), a BERT model pre-trained in the biomedical domain, specifically taking texts from PubMed. In this way, vectors of each biomedical entity are obtained, giving as input each of the biomedical concepts obtained from the graphs in a text format to the tokenizer.

In the architecture, the BERT Large pre-trained language model is used. This model will receive each of the statements indicated in the following section as input data.

4.4 Language Model Statements

The input to the language model associated with the node is the entire context of the text in question, such as Text 2.

$$[CLS] + Sentence + [SEP] + term + [SEP] + Label \quad (2)$$

Where [CLS] and [SEP] are the special classification and separation tokens used in BERT, respectively. Considering Text 1, we would have has input for term *pharmacological treatments*, labels *healthcare activity* and *research activity*, having two different inputs for the LLM.

The information obtained from the graphs after using the GNN proposed by Yasunaga et al. (2021) is combined with the output of the language model, representing that graph along with the contextual node obtained from the language model but adapted to the size of the GNN nodes. The pre-trained language model will return an embedding size equal to its last hidden layer.

4.5 Classification problem

The proposed classification problem will try to classify each entity detected in the target text among the 20 reduced categories obtained from the UMLS semantic types. The proposed architecture as in Yasunaga et al. (2021) employs an MLP at the end of the architecture. This MLP receives as input data the pooling vector obtained from the GNN, the output of each statement of the language model, and the vector that represents the context node Z. This concatenation will be received a total of 20 times, 1 for each category considered and will return a single probability that will be compared with the label in question.

The classification problem considered is multilabel, so each word to be classified can have more than one associated category, and in this case, no category is mutually exclusive. To carry out the classification, a sigmoid function and then binary cross entropy are used as the final activation function of the MLP, comparing each result obtained by the concatenation of a statement, graph, and context node vectors with the label in question.

The loss function considered is defined at the end of the MLP, so back-propagation updates the weights of the MLP and the GNN, as well as the linear transformations carried out to adjust the vectors representing each node of the graph to the dimensions of the GNN. The language model weights are kept frozen ($OntoLM_F$) or unfrozen ($OntoLM$) depending on the experiment.

5 *Experimentation*

The data obtained from the corpus are not correctly balanced, e.g. the category with the highest representation has 100 times more examples than the category with the lowest representation. This leads to performing an undersampling task on the data before training the model. Multi-Hop Graph Relation Network (MHGRN) introduced by Feng et al. (2020) also considers the number of different relations, but in previous question-answering experiments, the number of different relations is not large. In this case, the experiments consider all relations extracted from UMLS.

5.1 Undersampling

An undersampling task was carried out during the experiments to balance the dataset used. Balancing the dataset considerably improves the results obtained, since otherwise good results are only obtained with the labels with the highest representation. The final dataset used has a total of 800 instances for each of the labels, and each of these instances can have more than one label. During the undersam-

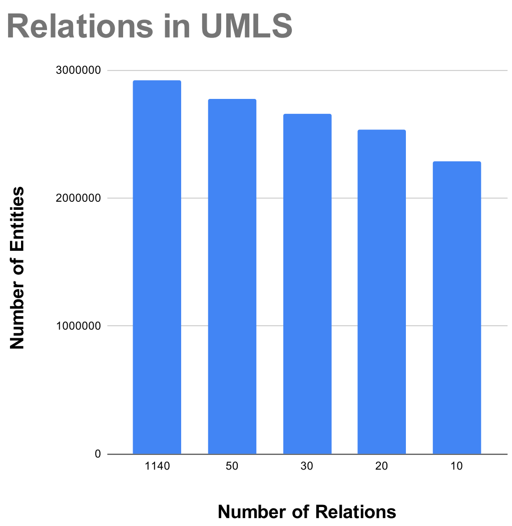


Figure 4: Number of different relations and the total coverage of entities using these relations, around 95 percent of the database uses only 50 relations.

pling task, training data was removed once the limit set per label was reached. Having a maximum of 800 in the most representative cases. Oversampling the data that has very few instances is not recommended since the training of the model gives very poor results on the categories where oversampling is performed. Therefore, removing categories with less than 800 instances is the best solution, reducing the dataset to 20 categories. The number of instances for each category is shown in Table 1.

5.2 Number of relations

All the relations present in the UMLS version have been used; however, of all these relations, a few have much higher representation reaching 95 per cent of the whole database downloaded with the top 50 relations, not counting the introduced relations *belongs to* and *meaning of*. With this in mind, a large part of the database can be represented with few relations, which is likely to positively affect the classification task by reducing the training complexity of the GNN. The representation of the database considering the number of relations can be seen in Figure 4. In this case, experiments using a simplified database with reduced relations will be conducted in future works.

5.3 Baseline

To carry out the experiment, the pre-trained BERT model is considered as baseline, specifically the large version obtained from the HuggingFace library together with an MLP comprising two hidden layers for final classification. This language model is considered the baseline since the whole system will use this model and the rest of the proposed architecture, to perform the classification.

6 Results

Precision and recall have been measured for each category considered during the experiment. Specifically, confusion matrices were used for each category, thus obtaining true positives, true negatives, false positives, and false negatives. In this way, the F1 score of each category was obtained, and the overall results can be seen in Table 2. Figure 5 shows the best micro results obtained for the model with better macro F1 ($OntoLM_F$).

Model	Accuracy	Precision	Recall	F1
Baseline	0.97	0.42	0.83	0.56
OntoLM	0.96	0.59	0.62	0.60
OntoLM$_F$	0.97	0.74	0.62	0.68

Table 2: Macro Accuracy, Precision, Recall and F1 for each experiment. Results for the best epoch.

7 Discussion

The proposed final architecture trains 1.2 million parameters, 300 times less than pre-trained language models such as BERT Large. However, the training becomes computationally expensive due to the large number of tensors representing graphs used as model input data compared to classical language model training, which employs only text tensors during this stage. The experiments were carried out using one 40 GB A100 GPU, spending a total of 12, 18 and 3 hours for training three epochs on *Baseline*, *OntoLM* and $OntoLM_F$, respectively and incrementing the batch size as much as possible to fill the GPU memory. Moreover, considering a graph and a statement for each possible term category increases the computational cost considerably. An attempt has been made to reduce the computational cost of the input data by reducing the size of the graphs, since in the case of the experiments carried out by Yasunaga et al. (2021), the size of the graphs used is 200

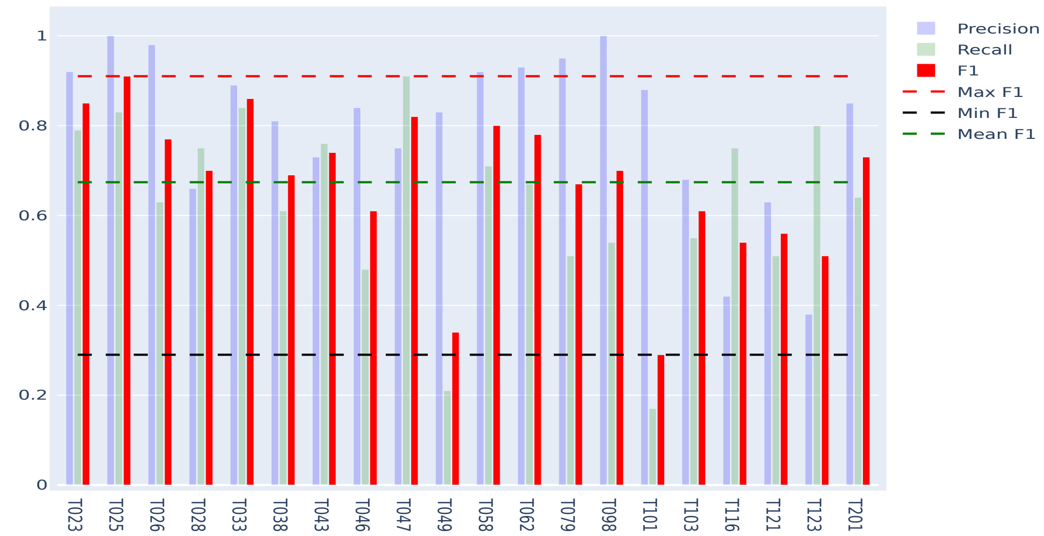


Figure 5: Precision, Recall and F1 for the best experiment in each of the 20 categories used.

nodes at most. In our particular case, reducing the size of the graphs to 100 nodes at most allows us to obtain good representations of each instance while reducing the weight of the input tensors by 40 per cent.

During the evaluation of the training epochs, notably during the different experiments, the real positives were not learned in the first two epochs. This is associated with the fact that this problem is a multilabel classification with too many negative categories, i.e., most of the label categories are zero, so returning zero in all categories for each training instance reduces the loss function considerably. Next, to reduce the loss function, it is necessary to identify the true positives in the output.

Table 2 shows that significant improvements have been obtained when using the proposed architecture over the baseline. However, the recall obtained in the baseline is far superior to the OntoLM experiments, suggesting that proper hyperparameter tuning is likely to give better results when running the full architecture. Running experiments with the full architecture yields better results by keeping the language model frozen ($OntoLM_F$), suggesting that the GNN architecture better adapts the knowledge of the language model for downstream tasks compared to the unfrozen language model ($OntoLM$). This result suggests that the proposed architecture can serve as an alternative to fine-tuning or that we can improve the results obtained by initially performing traditional fine-tuning on the language model and then attaching it to the overall architecture by training the GNN. As an alternative to fine-tuning, the proposal presented in this work is valid as, in addition to the better results, the computational cost (both in time and resources) is considerably reduced if the language model is kept frozen. The code of the experiments is available on GitHub [1].

During the realisation of each experiment, notably in the first two training epochs, the models do not classify any statement as positive, thereby obtaining only true and false negatives. The architecture finds as a first valid option to optimise all results in this way to reduce the loss function. Then, if the learning rate is low enough to classify the true positives, each model will learn to classify them, obtaining the best results in the first 10 epochs. This is quite likely considering that the labels used have very few positive categories, with 1 or 2 out of 20 in most cases.

Finally, the initial embeddings of each graph are not as expressive as they could be, mainly because the relations between the different nodes are not represented with contextualised embeddings from the beginning as with other methods. It is worth testing in future work by initialising these nodes with contextual embeddings based on their respective ontology and modifying the GNN architecture to process those contextualized embeddings.

[1]https://github.com/FabioDataGeek/OntoLM

8 Conclussion and Future Work

Given the results, multilabel classification tasks are improved by incorporating external structured knowledge. As far as we know, few works have performed the classification task with such a high number of categories. In the case of (Lee, Lee, and Ahn, 2022) they use 45 categories to perform multilabel classification of texts. However, in our case, the objective is not to classify the text but the possible entities found in a text from a certain domain. To the author's knowledge, very few works perform this specific task with so many categories. However, in tasks such as classification based on International Classification of Diseases codes, 10th edition (ICD-10), within the biomedical field (Gérardin et al., 2022) both entity and text classification studies exist, a task that is especially relevant to the purpose of this work.

Experiments show us an alternative way of adapting a language model to a specific domain without changing the domain weights, which is less computationally expensive and faster than loading the language models for fine-tuning. However, the time spent preprocessing the data to generate each graph must also be considered. The results obtained with the proposed architecture open up several lines of research, including the following:

1. The combination of ontologies with language models in other domains to perform classification tasks. Using this architecture with other ontologies can be especially useful to cover other NLP tasks such as word sense disambiguation with WordNet (Fellbaum, 1998).

2. Classification of texts with ICD-10 codes, using many categories and extending the experiment with ontological knowledge. UMLS is particularly interesting in this particular case, as it has specific information on ICD-10 codes.

3. Distillation of knowledge from language models, capturing the knowledge inside the language model using the GNN, with a final architecture much smaller than an LLM. If enough knowledge of the language model can be captured in the GNN, an architecture that detaches the language model can perform the same classification task.

4. Explainable and traceable NLP models from well-defined graph architectures and their respective GNN. After training the model, inference can be made with new data, and the activation of the different components of the GNN can be seen to determine the prediction obtained as suggested by (Ying et al., 2019).

5. Optimise the proposed architecture to avoid over-fitting while training the classifier with datasets similar to the proposed one and coupling previously fine-tuned language models.

6. Consider alternative training methods for classification with a large number of labels, in this case, modifying the loss function according to the category to be classified ((Su et al., 2022), (Hüllermeier et al., 2020).

Acknowledgments

This research has been funded by the University of Alicante, the Spanish Ministry of Science and Innovation, the Generalitat Valenciana, and the European Regional Development Fund (ERDF) through the following funding: At the national level, the following projects were granted: Coolang (PID2021-122263OB-C22); CORTEX (PID2021-123956OB-I00); *CLEARTEXT* (TED2021-130707B-I00); and SOCIALTRUST (PDC2022-133146-C22), funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by ERDF A way of making Europe, by the European Union or by the European Union NextGenerationEU/PRTR. At regional level, the Generalitat Valenciana (Conselleria d'Educacio, Investigacio, Cultura i Esport), granted funding for NL4DISMIS (CIPROM/2021/21).

References

Agarwal, C., O. Queen, H. Lakkaraju, and M. Zitnik. 2023. Evaluating explainability for graph neural networks.

AlKhamissi, B., M. Li, A. Celikyilmaz, M. Diab, and M. Ghazvininejad. 2022. A review on language models as knowledge bases.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the*

2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Bodenreider, O. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Chen, H., X. Liu, D. Yin, and J. Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 19(2):25–35, nov.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Elazar, Y., N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, and Y. Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Fellbaum, C., editor. 1998. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA.

Feng, Y., X. Chen, B. Y. Lin, P. Wang, J. Yan, and X. Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online, November. Association for Computational Linguistics.

Gehman, S., S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, November. Association for Computational Linguistics.

Gu, Y., R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1), oct.

Gérardin, C., P. Wajsbürt, P. Vaillant, A. Bellamine, F. Carrat, and X. Tannier. 2022. Multilabel classification of medical concepts for patient clinical profile identification. *Artificial Intelligence in Medicine*, 128:102311.

He, H., H. Zhang, and D. Roth. 2022. Rethinking with retrieval: Faithful large language model inference.

Huang, N., Y. R. Deshpande, Y. Liu, H. Alberts, K. Cho, C. Vania, and I. Calixto. 2022. Endowing language models with multimodal knowledge graph representations.

Hüllermeier, E., M. Wever, E. L. Mencia, J. Fürnkranz, and M. Rapp. 2020. A flexible class of dependence-aware multi-label loss functions.

Jiang, X., Y. Shen, Y. Wang, X. Jin, and X. Cheng. 2020. Bakgrastec: A background knowledge graph based method for short text classification. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, pages 360–366, Los Alamitos, CA, USA, aug. IEEE Computer Society.

Kaur, J., S. Bhatia, M. Aggarwal, R. Bansal, and B. Krishnamurthy. 2022. LM-CORE: Language models with contextually relevant external knowledge. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 750–769, Seattle, United States, July. Association for Computational Linguistics.

Lee, E., C. Lee, and S. Ahn. 2022. Comparative study of multiclass text classification in research proposals using pretrained language models. *Applied Sciences*, 12(9).

Li, Y., D. Tarlow, M. Brockschmidt, and R. Zemel. 2017. Gated graph sequence neural networks.

Liu, F., E. Shareghi, Z. Meng, M. Basaldella, and N. Collier. 2021. Self-alignment pretraining for biomedical entity representations.

McCray, A. 1989. The umls semantic network.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space.

Mrkšić, N., D. Ó Séaghdha, B. Thomson, M. Gašić, L. M. Rojas-Barahona, P.-H. Su, D. Vandyke, T.-H. Wen, and S. Young. 2016. Counter-fitting word vectors to linguistic constraints. In K. Knight, A. Nenkova, and O. Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, California, June. Association for Computational Linguistics.

Neumann, M., D. King, I. Beltagy, and W. Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, editors, *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August. Association for Computational Linguistics.

Peng, B., M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen, and J. Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Piad-Morffis, A., R. Muñoz, Y. Gutiérrez, Y. Almeida-Cruz, S. Estevez-Velarde, and A. Montoyo. 2019. A neural network component for knowledge-based semantic representations of text. In R. Mitkov and G. Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 904–911, Varna, Bulgaria, September. INCOMA Ltd.

Su, J., M. Zhu, A. Murtadha, S. Pan, B. Wen, and Y. Liu. 2022. Zlpr: A novel loss for multi-label classification.

Sun, J., C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, L. Ni, H.-Y. Shum, and J. Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations*.

Sun, Y., S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, W. Liu, Z. Wu, W. Gong, J. Liang, Z. Shang, P. Sun, W. Liu, X. Ouyang, D. Yu, H. Tian, H. Wu, and H. Wang. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation.

Wang, L., W. Zhao, Z. Wei, and J. Liu. 2022. Simkgc: Simple contrastive knowledge graph completion with pre-trained language models.

Wang, X., Q. He, J. Liang, and Y. Xiao. 2023. Language models as knowledge embeddings.

Yasunaga, M., H. Ren, A. Bosselut, P. Liang, and J. Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online, June. Association for Computational Linguistics.

Ying, Z., D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yáñez Romero, F., A. Montoyo, R. Muñoz, Y. Gutiérrez, and A. Suárez Cueto. 2023-09. A review in knowledge extraction from knowledge bases.

Zhang, Z., X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. 2019. ERNIE: Enhanced language representation with informative entities. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy, July. Association for Computational Linguistics.

Zhou, J., G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.

Tesis

Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 151-154 recibido 15-12-2023 revisado 20-01-2024 aceptado 27-01-2024

Automatic identification of Spanish academic collocations for an online writing tool

Identificación automática de colocaciones académicas en español para una herramienta en línea de ayuda a la redacción

Eleonora Guzzi
LyS Group, University of La Coruña
eleonora.guzzi@udc.es

Abstract: The present Ph.D. thesis was written by Eleonora Guzzi under the supervision of Dra. Margarita Alonso Ramos (University of La Coruña). The defense was held on the 11th of December 2023 at the University of La Coruña and the members of the committee were José Ignacio Pérez Pascual (University of La Coruña), the president, Moisés Almela Sánchez (University of Murcia), the secretary, and Amália Mendes (University of Lisbon), the vocal. The thesis was awarded an excellent grade Cum Laude, and the international doctorate mention.
Keywords: collocations, academic discourse, lexical complexity, corpus, writing tools

Resumen: La presente tesis doctoral ha sido escrita por Eleonora Guzzi, bajo la supervisión de la doctora Margarita Alonso Ramos (Universidad de la Coruña). La defensa se celebró el 11 de diciembre de 2023, en la Universidad de la Coruña, ante el tribunal formado por el presidente, José Ignacio Pérez Pascual (Universidad de la Coruña), el secretario, Moisés Almela Sánchez (Universidad de Murcia) y la vocal, Amália Mendes (Universidad de Lisboa). La tesis obtuvo la calificación de sobresaliente Cum Laude y la mención internacional.
Palabras clave: colocaciones, discurso académico, complejidad léxica, corpus, herramientas de escritura

1 Goals and motivation

The main goal of the present thesis is to contribute to the field of Spanish for Academic Purposes through a research on academic collocations within the areas of Corpus Linguistics, Computational Linguistics and Lexicography.

On the one hand, the choice of academic discourse as the scope of this study has been motivated by the fact that one of the greatest difficulties faced by students is to acquire proficiency in academic writing. This challenge arises especially because an advanced knowledge of specific vocabulary is required, among other genre-related knowledge. On the other hand, the decision to focus on collocations, such as *deep analysis*, *confirm hypothesis* or *draw conclusions,* as object of study has been based on the assumption that this type of expressions, along with other phraseological units, enriches the academic prose and contributes to communicative effectiveness. Moreover, research has proved that a higher collocational competence in academic discourse can be synonymous with greater academic success. However, students typically exhibit limited experience in academic writing and have a little exposure to this type of vocabulary: academic collocations are neither part of implicit everyday language nor explicitly taught as the technical vocabulary. In fact, existing literature has highlighted the insufficient familiarity undergraduate students possess with the prototypical phraseology of academic discourse (Boers & Webb, 2018; Paquot & Granger, 2012) and identified collocations as a persistent and frequent challenge in the written competence of both second language learners and less proficient native speakers. In addition, if we explore the field of Spanish academic writing, scarce studies focus on the identification of academic vocabulary and lexicographic resources related

ISSN 1135-5948 DOI 10.26342/2024-72-11

to phraseological expressions and academic writing.

Therefore, this research aims to address identified gaps in quantitative corpus-based studies on Spanish academic vocabulary by compiling a list of academic collocations and providing a comparison of the collocational complexity of expert and undergraduate writings. Moreover, given the scarcity of resources and lexicographic tools in Spanish as far as academic writing is concerned (Alonso-Ramos et al., 2017; Núñez Cortés & Da Cunha, 2022; Guzzi & Alonso-Ramos, 2023a), one of the purposes of this research is precisely to include the Spanish academic collocations list in a writing aid (HARTA; http://www.dicesp.com:8083; Alonso-Ramos et al. 2017), designed with a corpus-dictionary format and aimed to contain academic multi-word expressions in Spanish (Guzzi et al. 2023).

On the other hand, as a result of the thresholds established in the collocational complexity comparison, an automated evaluation system intended for academic Spanish certification exams, such as EXELEEA (Mendoza, 2015) is proposed. Finally, the study of collocations could have multiple practical applications, which in Spanish have not yet been exploited. Thus, the proposal of this thesis could be applied as a didactic resource for the teaching of academic Spanish in writing centers, as well as for students with Spanish as L2 who access Spanish-speaking universities. It could have also applications in the field of Computational Linguistics and Natural Language Processing: as is well known, collocational resources are especially useful for rule-based generation and translation systems.

2 *Outline of the dissertation*

This thesis comprises eight chapters. Chapter 1 provides an overview of the research's motivation and main objectives. In Chapter 2, the field of academic discourse is introduced, encompassing a review of approaches, to delineate the study's scope and an explanation of key concepts such as Language for Specific Purposes, specialized languages and scientific discourse. In this chapter, a general description of academic vocabulary is presented, as well as the types of phraseological units approached within this type of discourse, that includes the concept of collocation adopted in this study (Mel'čuk, 2015). Automatic, statistical and manual methods for identifying academic words and collocations are outlined through a comprehensive review of existing vocabulary lists. The final section explores the concept of lexical and phraseological complexity related to vocabulary assessment, involving parameters such as diversity and sophistication (Crossley, 2020; Kyle & Crossley, 2015; Paquot, 2019). Chapter 3 delves into academic tools addressing phraseological units, ranging from academic corpora to online writing aids. In addition, a first version of HARTA is presented in detail together with a usability test of the tool. Finally, the resources focused on vocabulary assessment related to lexical complexity and lexical profile are addressed.

Chapter 4 details the composition and processing of the two corpora that are employed in this research (expert and novice), using NLP techniques, as well as the methodology for automatically extracting collocation candidates. Chapter 5 explains exhaustively the compilation of the list of Spanish academic collocations. Criteria for filtering collocations are presented, including phraseological and interdisciplinary statistical criteria, along with the method followed to validate the list. Chapter 6 contrasts expert and novice use of collocations, by means of collocational complexity of their texts, that includes the parameters of sophistication, known as the property of lexical items that are less common in general language and are more formal or typical of academic discourse, and diversity, known as the index of repetition of lexical items.

Chapter 7 explores practical applications of the Spanish academic collocation list, emphasizing its integration into the HARTA tool, together with quantitative data and improvements based on the results of the usability test. The second part introduces a beta version of an evaluation tool to automatically calculate the collocational profile of texts, aimed for teachers, evaluators, and researchers. Finally, Chapter 8 summarizes the main conclusions, acknowledges limitations, and outlines future directions for research.

3 *Main contributions*

From this research, four main contributions can be retrieved.

3.1. Spanish Academic Collocations List

The first one has been the development of a reference list of 5.402 Spanish academic collocations based on a large expert corpus, consisting of scientific articles from 12 different disciplines.

This includes a detailed description of the procedure and criteria we followed to identify these collocations: an automatic extraction process and a manual and statistical review that includes phraseological and distribution criteria. The phraseological criteria allowed us to discard either free combinations or idioms. This phase has proved to be the most demanding due to the large number of candidates automatically extracted and because sometimes the boundaries between types of combinations can be blurred. Even this selection process, some specialized collocations still persisted in the selection, that highlighted the need of an interdisciplinarity analysis. It involved distribution filters and the identification of the different senses of the collocations' bases to discard specialized units.

Furthermore, a comparison of the collocations obtained following this method is compared to the collocations that would have been obtained if two association measures (log-likelihood and Mutual Information) were used above a specific threshold. This comparison has shown that the overlapping degree is not elevated but that the log-likelihood measure could be better than Mutual Information for the identification of Spanish academic collocations with a lower threshold.

3.2. Collocational profile of expert and novice writings

The second result concerns the contrastive analysis of the use of academic collocations by experts and novice writers by means of collocational complexity (Guzzi & Alonso-Ramos, 2023b). Results have shown that experts have a wider repertoire of collocations and use those that are stylistically more salient in academic discourse. However, results also suggested that, sometimes, they repeat the same collocation several times in the same text. On the other hand, results have indicated that scientific areas may influence the score of collocational complexity: Biology and Health Science is the field in both groups (expert and novices) where more sophisticated collocations are used, but Social Sciences shows the highest amount of academic collocations. Moreover, a correlation between the linguistic general quality of text and level of collocational complexity was corroborated. Finally, the results obtained allowed us to establish a threshold for scoring academic texts according to the number of collocations, diversity and sophistication. For this purpose, we applied a series of formulas relating to collocational diversity and sophistication in order to obtain the collocational profile of the texts analyzed, understood as an image reflecting the collocational competence of academic texts' writers.

3.3. Improvements of HARTA

The third result is related to the integration of collocations in the writing aid tool HARTA, with improvements in their accessibility, functionality and amount of information. Data about the frequency and distribution of academic collocations has been integrated with a clearer view, as shown in Figure 1, as well as collocations associated to two possible meanings. Those improvements have been implemented as a result of the establishment of the collocation list and the usability test of the tool.

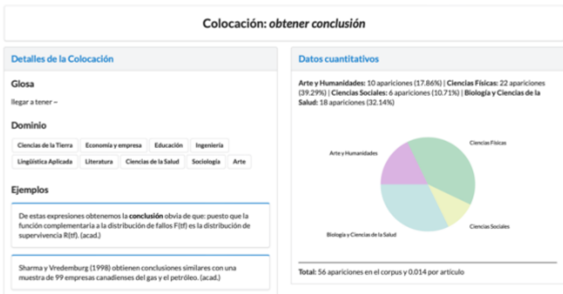


Figure 1. Lexicographic entry of the academic collocation *obtain conclusion* in HARTA.

3.4. Evaluation tool for the collocational complexity of academic texts

The fourth and last contribution has been the proposal for an automatic evaluation system of the collocational competence of writers through the analysis of collocational complexity. Using a series of Python scripts, a text is processed and a set of indexes from the text are identified that includes: the number of words; the collocational lemmas and their frequency; and the collocations diversity and sophistication, along with a global score, as shown in Figure 2.

Figure 2. Part of the results that the evaluation tool on collocational complexity shows when a text is analyzed.

Acknowledgments

This research was funded by the following grants, centers and institutions: Ministerio de Ciencia e Innovación (PID2019-109683GB-C21); Programa de Axudas á Etapa predoutoral da Xunta de Galicia, FSE Galicia 2014-2020; Centro de Investigación de Galicia do Sistema Universitario de Galicia "CITIC", funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

References

Alonso-Ramos, M., M. García-Salido, and M. Garcia. 2017. Exploiting a corpus to compile a lexical resource for academic writing: Spanish lexical combinations. In Iztok Kosem, Jelena Kallas, Carole Tiberius, Simon Krek, Miloš Jakubíček and Vít Baisa (eds.), *Electronic lexicography in the 21st century, Proceedings of 2017 eLex Conference* (pages 571-586). Brno (Czech Republic).

Boers, F. and S. Webb. 2018. Teaching and learning collocation in adult second and foreign language learning. *Language Teaching*, 51:1, 77-89.

Crossley, S. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11:3, 415-443.

Guzzi, E. and M. Alonso-Ramos. 2022. Selección de colocaciones académicas en español a través de un filtro de interdisciplinariedad. *Procesamiento del Lenguaje Natural*, 69, 83-94.

Guzzi, E. and M. Alonso-Ramos. 2023a. Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, 2, 1-22.

Guzzi, E. and M. Alonso-Ramos. 2023b. Sofisticación y diversidad como medidas de complejidad léxica para determinar el perfil colocacional de textos académicos en español. *Revista Signos*, 56:112.

Guzzi, E., M. Alonso-Ramos, M. Garcia, and M. García-Salido. 2023. Annotation of lexical bundles with discourse functions in a Spanish academic corpus. In Archna Bhatia, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, y Shiva Taslimipoor (Eds.), *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)* (pages 99-105), Association for Computational Linguistics. Dubrovnik (Croatia).

Kyle, K. and S. Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, 49:4, 757-786.

Mel'čuk, I. 2015. Clichés, an understudied subclass of phrasemes. *Yearbook of Phraseology*, 6:1, 55-86.

Mendoza, A. 2015. La validez en los exámenes de alto impacto: Un enfoque desde la lógica argumentativa. *Perfiles educativos,* 37:149, 169-186.

Paquot, M. and S. Granger. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, 32, 130-149.

Paquot, M. 2019. The phraseological dimension in interlanguage complexity research. *Second language research,* 35:1, 121-145.

Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 155-158 recibido 15-12-2023 revisado 20-01-2024 aceptado 27-01-2024

Analysis and classification of spam email using Artificial Intelligence to identify cyberthreats

Análisis y clasificación de correo electrónico no deseado mediante Inteligencia Artificial para la identificación de ciberamenazas

Francisco Jáñez Martino
Departamento de Ingeniería Eléctrica y de Sistemas y Automática, Universidad de León
Campus de Vegazana, s/n, 24007 León, España
francisco.janez@unileon.es

Abstract: Summary of the Ph.D. thesis written by Francisco Jáñez Martino and supervised by Prof. Dra. Rocío Alaiz Rodríguez and Dr. Víctor González Castro at Universidad de León. The defense of the thesis was in León (Spain) in 21st of December 2023 by a committee formed by Dr. Arturo Montejo Ráez (Universidad de Jaén, Spain), Dr. Petr Motlicek (Idiap Research Institute, Switzerland), and Dra. Laura Fernández Robles (Universidad de León, Spain). An international mention was garnered following a six-month tenure at the Universitá di Bologna under the supervision of Dr. Alberto Barrón Cedeño. This Ph.D. thesis was awarded an outstanding Cum Laude grade.
Keywords: Spam email classification, Machine Learning, Attention models, Natural Language Processing, Persuasion detection, Risk classification, Cybersecurity

Resumen: Tesis doctoral realizada por Francisco Jáñez Martino y supervisada por la Prof. Dra. Rocío Alaiz Rodríguez y el Dr. Víctor González Castro en la Universidad de León. La defensa de la tesis se realizó en León (España) el 21 de diciembre de 2023 ante un tribunal compuesto por el Dr. Arturo Montejo Ráez (Universidad de Jaén, España), el Dr. Petr Motlicek (Idiap Research Institute, Suiza), y la Dra. Laura Fernández Robles (Universidad de León, España). Se obtuvo la mención internacional tras una estancia de 6 meses en la Università di Bologna bajo la supervisión del Dr. Alberto Barrón Cedeño. La tesis obtuvo una calificación de sobresaliente Cum Laude.
Palabras clave: Clasificación de correos spam, Aprendizaje Automático, Modelos de atención, Procesamiento del Lenguaje Natural, Detección de la persuasión, Predicción del riesgo, Ciberseguridad

1 *Introduction*

Spam email has been a problem since the creation of this popular communication medium. Traditionally, these unwanted and unsolicited emails contained advertisements, strange chains or just annoying messages. Due to the rise of Internet and electronic devices, cybercriminals leverage the accessibility of free payment, anonymity and massive use of email services to spread malware, phishing or spoofing attacks among other scams. This turns spam into a big data problem as well as a current cybersecurity challenge.

The main solution for detecting spam email are the anti-spam filters, which showed high performance in the literature. These filters are currently based on Natural Language Processing (NLP) and Machine Learning (ML) models (Dada et al., 2019). However, users still report attacks rooted in spam emails. Hence, understanding, analysing and classifying how spammers design these emails has become a mandatory stage, not only to enhance filtering but also to improve the extraction of information.

In this Thesis, we introduced novel models, methodologies, approaches, and datasets for the analysis and identification of emerging cybersecurity threats in spam

ISSN 1135-5948 DOI 10.26342/2024-72-12

emails. Motivated by our collaboration with the Spanish National Institute of Cybersecurity (INCIBE), our dedication lies in creating applications and conducting research to enhance the early detection of risky and malicious emails. Our approach heavily relies on the application of NLP, as well as Machine and Deep Learning techniques, mainly centred around supervised learning methods.

Several contributions outlined in this dissertation are intended to be integrated into tools being developed by Law Enforcement Agencies (LEAs) and INCIBE. These tools aim to provide more comprehensive and timely alerts to organizations and citizens regarding potential risks posed by spam email. This thesis proposed models aimed at ensuring the security, integrity, and privacy of users in the face of cyberattacks originating from spam emails.

Our main objectives were: a) classifying spam emails according to their cybersecurity topic, b) spotting both the presence of persuasion and the specific techniques employed and c) extracting potentially useful information from both their headers and body to spot risky emails. Additionally, many of the data mining and NLP techniques can be utilized for similar issues, such as smishing, fraudulent content on websites, or social media.

2 Thesis Overview

This thesis consists of seven chapters, which are described as follows:

Chapter 1 We outlined the objectives and motivation behind the thesis. Our motivation moving away from the traditional spam filtering to provide support for cybersecurity organizations to comprehend the properties of spam emails and present models to expedite and enhance their analysis.

Chapter 2 We reviewed the state-of-the-art anti-spam filters, and found that they showed high performance on outdated datasets during their evaluation. However, their assessment did not consider two challenging problems in the spam domain: dataset shift and spammer strategies to deceive these filters. Our review encompassed the investigation of dataset shift in ML models considering adversarial environments and works related to detecting specific spammer strategies. In depth, we reviewed the study of spammer tricks like obfuscated words, poisoning text, hidden text, image-based spam and other emerging trends. Moreover, we carried out an empirical experimentation to provide supporting evidences. Finally, we explored the existing cybersecurity challenges associated with spam email. The review and experimentation of this Chapter has been published in Artificial Intelligence Review journal (Jáñez-Martino et al., 2022).

Chapter 3 In this Chapter, we addressed the development of a text classifier capable of identifying the cybersecurity topic of a spam email. We used a hierarchical clustering and manual inspection to define eleven cybersecurity classes for the first time in the literature.

We conducted an evaluation (per language) of the combinations of two traditional approaches, Term Frequency - Inverse Document Frequency (TF-IDF) and Bag of Words (BoW) and two word embeddings models, word2vect and BERT (Devlin et al., 2018), as text representation techniques along with four popular ML classifiers: Support Vector Machine (SVM), Naïve Bayes, Logistic Regression and Random Forest.

We also provided the confusion matrices, an evaluation of the models performance per class and a data augmentation analysis using both reducing the majority classes and increasing the minority class. The work of this Chapter has been published in Applied Soft Computing journal (Jáñez-Martino et al., 2023). A preliminary study was published on Arxiv (Jáñez-Martino et al., 2020).

Chapter 4 In the fourth Chapter, we sought to identify persuasive elements in spam email. Upon reviewing the state of the art, we found out theoretical and psychological works associated with different kinds of spam emails like phishing emails. Due to this fact, we started from relating the persuasive principles presented in (Ferreira, Coventry, and Lenzini, 2015) to the growing attention in developing automatic models to spot persuasion and those techniques involved in news articles (Da San Martino et al., 2019). Theses works set the basis of our hypothesis, which is to analyze the role of persuasion in manipulating users to perform an specific action such as clicking in an external link or opening an attachment.

We designed NLP models at three levels of granularity: full email, sentences and span text (one or more words but always shorter than a sentence). We detailed our approach to use the datasets and models derived from persuasion in news articles for full email and span text classification. For sentence classification, we described the creation of a manually annotated dataset following binary and multilabel annotations and adjusted pre-trained models.

This chapter has been covered by a paper presenting the whole study and submitted to a journal.

Chapter 5 In this Chapter, we aimed to extract further information from spam email to improve the warnings launched by cybersecurity agencies to report organization and citizens about spam campaigns and frauds involving harmful and risky emails (Gallo et al., 2021). We followed two approaches: a) binary classification (high and low risk) and regression (scaling the email in a level of risk range from 1 to 10).

We analyzed the spam email through a NLP feature extraction according to reported key points pointed out by cybersecurity experts. We also used the previous cybersecurity topics as features and conducted an extend investigation to determine the quality of email address of spammer senders. We explained every feature and analyzed the relevance of each one and group for classification.

The extended work on address classification was presented at the Document Engineering 2021 conference (Jáñez Martino et al., 2021). The paper presenting the whole system has been submitted to a journal.

Chapter 6 We highlighted our eight main findings and future work, emphasizing the expansion of some research lines. In addition, we expressed our interest in applying the methodologies developed in this thesis to other domains, such as social media or instant messaging.

Chapter 7 In compliance with university requirements, we translated the conclusions and future work presented in Chapter 6 into Spanish.

3 Contributions

We enumerated the principal contributions of this thesis as follows:

We outlined the spam filtering, spammer strategies and dataset shift problem. We empirically demonstrated how these factors negatively impact the evaluation of anti-spam filters. We compared the performance of filters when being trained on one dataset and evaluated on other dataset (using five of the most used spam emails datasets, both back and forth in time). This review underscored the importance of comprehending spam properties to enhance both the filters and their assessment. Additionally, it also highlighted the need to study the detection of spammer strategies such as hidden text, word obfuscation, or text embedded in images, as well as other emerging tricks like mixing languages.

We semi-automatically labeled a novel dataset in the spam emails domain by using hierarchical clustering and visual inspection through a collection of INCIBE spam emails. Our dataset called Spam Email Classification (SPEMC) holds almost 15k spam emails per language (both English and Spanish) divided into eleven cybersecurity topics. These are academic media, extortion hacking, fake reward, health, identity fraud, money making, pharmacy, service, sexual content dating, work offer and other.

We presented a text classification pipeline based on traditional text representation techniques and word embeddings along with four popular ML algorithms to detect the eleven cybersecurity topics. This pipeline includes an email processing stage to extract all textual content from the subject, body and images. We considered the appearance of spammer strategies, such as image-based message or hidden text, and we applied Optical Character Recognition (OCR) techniques to extract only the visible text.

We developed automatic systems to detect persuasion and its techniques at different levels of granularity: full email, sentences and text spans. We replicated ML models based on NLP features as well as fine tuning pretrained attention models. For sentences classification, we manually annotated sentences of spam emails based on binary, persuasive or not, and multilabel perspective, containing eight persuasion techniques labels plus the negative one.

We introduced a novel set of 56 features based on NLP to discriminate those spam emails with more potential risk for individuals and organizations. We divided the features into five groups: headers, text, attachments, URLs and protocols. We developed models following two approaches: classification and regression.

We manually annotated two spam email datasets collected in different sources, one private from INCIBE resources and one public from Spam Archive of Bruce Guenter[1], based on their potential risk. We labeled them for (i) a regression problem using a scale of risk (1-10) and (ii) a classification problem distinguishing two classes, low and high risk. Low risk spam refers to messages that closely resemble traditional ones containing advertisements and annoying content, but without the presence of malware or scams that could end exposing leaked data of users. While high risk level include cybersecurity attacks such as spreading ransomware, phishing, spoofing or extortion.

We evaluated three classifier and three estimators using our novel set of features as input. We conducted an analysis of feature importance for the classification approach by systematically removing or retaining one set of features. We also evaluated the relevance of removing each individual feature one by one to establish a cutoff number of features. Due to the high relevance of the address classification according to cybersecurity experts, we carried out an extended investigation. The objective was to classify the address of a spam sender into low and high quality. To do this, for the first time in the literature, we presented a set of 18 features extracting information from the username, domain and top-level domain (TLD) of each address and fed up four ML classifiers for evaluation using a manually labeled dataset call Email Address Quality - 6k. A high quality is given when the address contains popular brands or email services, truthful TLDs and imitate common user's address without random number, characters or letters and short username or domains.

Acknowledgements

This work was supported by the framework agreement between the Universidad de León and INCIBE (Spanish National Cybersecurity Institute) under Addendum 01 and the Predoctoral Grant of Junta de Castilla y León.

References

Da San Martino, G., S. Yu, A. Barrón-Cedeño, R. Petrov, and P. Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646.

Dada, E. G., J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa. 2019. Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon*, 5(6):e01802.

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805:1–16.

Ferreira, A., L. Coventry, and G. Lenzini. 2015. Principles of persuasion in social engineering and their use in phishing. In T. Tryfonas and I. Askoxylakis, editors, *Human Aspects of Information Security, Privacy, and Trust*, pages 36–47, Cham. Springer International Publishing.

Gallo, L., A. Maiello, A. Botta, and G. Ventre. 2021. 2 years in the anti-phishing group of a large company. *Computers & Security*, 105:102259.

Jáñez Martino, F., R. Alaiz-Rodríguez, V. González-Castro, and E. Fidalgo. 2021. Trustworthiness of spam email addresses using machine learning. In *Proceedings of the 21st ACM Symposium on Document Engineering*, DocEng '21, page 4, New York, NY, USA. Association for Computing Machinery.

Jáñez-Martino, F., R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre. 2022. A review of spam email detection: analysis of spammer strategies and the dataset shift problem. *Artificial Intelligence Review*, 56:1145–1173.

Jáñez-Martino, F., R. Alaiz-Rodríguez, V. González-Castro, E. Fidalgo, and E. Alegre. 2023. Classifying spam emails using agglomerative hierarchical clustering and a topic-based approach. *Applied Soft Computing*, 139:110226.

Jáñez-Martino, F., E. Fidalgo, S. González-Martínez, and J. Velasco-Mata. 2020. Classification of spam emails through hierarchical clustering and supervised learning.

[1]http://untroubled.org/spam/ retrieved December 2023

Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 159-164 recibido 15-12-2023 revisado 20-01-2024 aceptado 27-01-2024

Detecting offensive language by integrating multiple linguistic phenomena

Detección del lenguaje ofensivo mediante la integración de diferentes fenómenos lingüísticos

Flor Miriam Plaza-del-Arco
[1]Bocconi University, Milan, Italy
flor.plaza@unibocconi.it

Abstract: This is a summary of the Ph.D. thesis conducted by Flor Miriam Plaza del Arco at the University of Jaén under the supervision of Ph.D. M. Teresa Martín Valdivia and Ph.D. L. Alfonso Ureña López. The thesis defense took place in Jaén on January 30, 2023, with the doctoral committee comprising Ph.D. Mariona Taulé Delor from the University of Barcelona, Ph.D. José Camacho Collados from the University of Cardiff, and Ph.D. Eugenio Martínez Cámara from the University of Granada. Notably, the thesis was awarded the distinction of Summa Cum Laude and received international recognition.
Keywords: Natural language processing, Spanish corpora, offensive language detection, hate speech, multitask learning, linguistic phenomena

Resumen: Este es un resumen de la tesis doctoral realizada por Flor Miriam Plaza del Arco en la Universidad de Jaén, bajo la supervisión de la Dra. M. Teresa Martín Valdivia y el Dr. L. Alfonso Ureña López. La defensa de la tesis tuvo lugar en Jaén el 30 de enero de 2023 y la comisión de doctorado estuvo formada por la Dra. Mariona Taulé Delor de la Universidad de Barcelona, el Dr. José Camacho Collados de la Universidad de Cardiff y el Dr. Eugenio Martínez Cámara de la Universidad de Granada. Cabe destacar que la tesis obtuvo la calificación de Summa Cum Laude y la mención internacional.
Palabras clave: Procesamiento del Lenguaje Natural, corpus en español, detección del lenguaje ofensivo, discurso de odio, aprendizaje multitarea, fenómenos lingüísticos

1 Introduction

One of the characteristics that distinguish humans from other living beings is the ability to communicate systematically and understandably, i.e. through language. Language is defined as a sophisticated system of both phonetic and written symbols that allows two or more individuals to communicate ideas, thoughts, sentiments, attitudes, and different situations. Since the emergence of Web 2.0, users were no longer limited to face-to-face communication but rather used online platforms to interact. This interaction has resulted in an increasing amount of textual data being available on the Web. Natural Language Processing, a tract of Artificial Intelligence and Linguistics, arises for the development of computational systems to interpret human language and thus enable human-computer interaction. Giving computers this skill offers a plethora of benefits, including the potential to moderate harmful conduct on social media.

This doctoral thesis focuses on both the creation of linguistic resources and the development of NLP-based techniques to aid in the automatic detection of offensive language on the Web. On the one hand, for the development of these techniques, data labeled are essential to learning the language patterns characteristic of this behavior; however, the available resources are mainly focused on English, leaving aside other languages such as Spanish with very scarce or non-existent resources of this nature. Therefore, a fundamental part of this doctoral thesis is focused on the generation of these resources for Spanish. On the

ISSN 1135-5948 DOI 10.26342/2024-72-13

other hand, for the implementation of automatic systems based on NLP, one of the main contributions has been the integration of different linguistic phenomena that might be involved in the expression of offensiveness in computational systems. In particular, we developed a Multitask Learning (MTL) method based on Transfer Learning (TL). We believe that this methodology plays an important role in their application to the detection of more specific problems in our society, such as Hate Speech (HS), misogyny, or sexism, that have been addressed in the frame of this doctoral thesis. As a result, it should be mentioned that this thesis has both a social and technological dimension to contribute to society's improvement.

1.1 Motivation

Social media have grown into the primary means of communicating between people, allowing users to have conversations, share opinions, and create content. The rise in digital social connections has led to the dissemination of harmful communication, which is sometimes aided by the anonymity afforded by these platforms (Aguilera-Carnerero and Azeez, 2016). As a consequence, offensive language and one of its most damaging forms, HS, tends to proliferate swiftly and is difficult to regulate. For instance, according to a Spanish report in 2020 on the evolution of hate crimes in Spain[1], threats, insults, and discrimination are counted as the most repeated criminal acts, with the Internet (45%) and social media (22.8%) as the most widely used media to commit these actions. Similarly, a recent survey on hate crimes in Spain 2021[2] shows that 41.65% of the participants, out of a total of 437, have been victims of hate crimes on more than one occasion in the last 5 years. On the one hand, they have received offensive comments on more than 10 occasions. On the other hand, more than 50% of them have received offenses or threats through social networks or the Internet. Finally, more than 70% of the respondents have received discriminatory treatment on one or more occasions in the last 5 years.

In this regard, inaction against offensive language allows for the further reinforcement of prejudices and stereotypes, while this type of hostile communication may lead to negative psychological effects among online users, causing anxiety, harassment, and, in extreme cases, suicide (Hinduja and Patchin, 2010). As a result, this scenario has motivated interested stakeholders (governments, online communities, and social media platforms) to look for efficient solutions to prevent Internet hostility. One strategy used to tackle this problem is through legislation, by implementing laws and policies. For instance, since 2013 the Council of Europe has sponsored the "No Hate Speech" movement[3] seeking to mobilize young people to combat HS and promote human rights online. In May 2016, the European Commission reached an agreement with Facebook, Microsoft, Twitter, and YouTube to implement the "Code of Conduct on countering illegal HS online"[4]. From 2018 to 2020, platforms such as Instagram, Snapchat, and TikTok adopted the Code. One of the initial and most common approaches to hatred intervention adopted by social media platforms is content moderation. This approach is based on the suspension of user accounts and the removal of hate messages while attempting to balance the right to freedom of expression.

Although these approaches have the clear advantage of analyzing the context and accurately identifying this behavior, still these strategies do not seem to achieve the desired effect because they involve an intense, time-consuming, and costly procedure that limits scalability and quick solutions. At the same time, hate content is continuously growing and adapting, making it harder to identify (Davidson et al., 2017). As a result of these challenges, an alternative and preferable option is to rely on NLP-based methods to automatically detect this type of harmful online communication. Advances in NLP can be used to detect offensive content online thus decreasing the time and effort in fighting this problem. Offensive language detection and analysis has become a major area of research in NLP. However, existing NLP-based methods face several drawbacks. Firstly, detecting offensive content is challenging for machines (Zampieri et al., 2019; Wiegand, Ruppenhofer, and Kleinbauer, 2019; Poletto et al., 2020), since this type of language presents a subjective nature as well as social and cultural implications. Though recent

[1] `https://shorturl.at/hlnAX`
[2] `https://shorturl.at/mpxLR`
[3] `https://shorturl.at/DQ345`
[4] `https://shorturl.at/kvHOT`

approaches of sequence-to-sequence models (Zampieri et al., 2020; Tontodimamma et al., 2021) have achieved good performance in detecting this type of content, most of them have not considered linguistic phenomena that may occur in the expression of offensive language such as those of an implicit nature such as sarcasm and irony (Chauhan et al., 2020; Wiegand, Ruppenhofer, and Eder, 2021). Secondly, since most of the available corpora contain messages from the Twitter platform, automatic systems have specialized in learning the language style and register used by the users on this platform, making cross-domain transfer difficult when using such systems on other platforms. Thirdly, so far most of the research to solve this problem has been focused on English (Fortuna and Nunes, 2018), leaving other languages such as Spanish in second place, although combating this type of behavior is a global concern.

These challenges motivate this doctoral thesis to explore methods for accurately detecting offensive language on the Web using NLP techniques to aid in this process. **This thesis relies on advanced methods in NLP such as deep learning to tackle this issue**. First, it faces the problem of limited training data, especially in Spanish, generating appropriate resources to combat offensive textual content. These resources will also help to solve the limitation of the systems specialized in Twitter since messages from other social platforms such as YouTube and Instagram are considered. Secondly, it introduces different linguistic phenomena that could be involved in the expression of offensiveness and could help in the detection of this content. Then, a novel method is proposed where these identified phenomena are integrated for the detection of offensive language, using state-of-the-art techniques based on transfer learning. Finally, this novel method is applied for the detection of different offensive language scenarios (HS, sexism, toxicity), analyzing which specific linguistic phenomena are beneficial in each of them.

1.2 Hypotheses

This thesis studies the problem of automatically detecting offensive textual language with deep learning techniques for NLP. The main hypothesis of this thesis is the following: **Advanced NLP methods based on deep learning, in particular transfer learning, aid in the detection of offensive textual language. We subdivide this hypothesis into three hypotheses:**

Hypothesis 1 (H1) The subjective nature of offensive language can have strong cultural, demographic, and social implications, and therefore language-specific resources and models are required.

Hypothesis 2 (H2) Transfer learning models leveraging linguistic phenomena related to offensive language expression outperform those that do not integrate this information in offensive language detection tasks.

Hypothesis 3 (H3) Incorporating specific linguistic phenomena into transfer learning methodologies can enhance the detection of various offensive scenarios. Offensive language detection encompasses a range of scenarios, such as identifying sexist content, hate speech, or toxic language.

2 Thesis outline

This thesis is structured into 8 chapters, outlined as follows:

- **Chapter 2** includes an overview of the background information that is significant for understanding the content of this thesis. We review traditional ML and Neural Network (NN) based methods for offensive language research in NLP. We furthermore provide a compilation of different existing resources labeled with offensiveness. Then, we present the research challenges and opportunities based on the previous research approaches reviewed.

- **Chapter 3** introduces our preliminary research in the thesis, focusing mainly on traditional ML approaches to address HS detection, including misogyny and xenophobia. In addition, we present the first experiments with monolingual and multilingual pre-trained language models based on Transformers.

- **Chapter 4** describes the different corpora and lexicons we generate during the thesis for the research on offensive language and emotion analysis. Specifically, three corpora and three lexicons, mainly focused on Spanish, are presented.

- **Chapter 5** introduces our contribution to addressing offensive language detection. We propose a novel approach that uses the MTL paradigm to combine different phenomena inextricably related to the expression of offensive language. This approach aims to benefit from shared knowledge across tasks to improve the detection of offensive language. We identify some linguistic phenomena that might be involved in the expression of offensive language and present initial experiments.

- **Chapter 6** focuses on the evaluation of the proposed MTL learning approach in different offensive language scenarios studying the integration of the linguistic phenomena defined in Chapter 5. We show the success of our MTL methodology by comparing its performance with previous state-of-the-art approaches that do not consider this useful information.

- **Chapter 7** presents two different shared tasks organized in the framework of this doctoral thesis to promote the research on emotion analysis and offensive language detection in Spanish. The task descriptions, the corpora and evaluation measures used as well as the participants and results achieved are described.

- **Chapter 8** finally summarizes our conclusions where we present the main findings of this doctoral thesis and suggest future research directions within offensive language research.

3 Main contributions

The research conducted in this doctoral thesis has resulted in several contributions that support the hypothesis outlined in Section 1.2.

Contributions to support H1:

Contribution 1 The generation of different linguistic resources for offensive language research and emotion analysis focused mainly on Spanish (Plaza-del-Arco et al., 2020; Plaza-del-Arco et al., 2021; Plaza-del-Arco et al., 2022).

Contribution 2 We have developed our annotation scheme for each of the resources generated.

Contribution 3 Using the resources generated, we have organized different shared tasks in the IberLEF evaluation campaign to promote offensive language research in Spanish (Plaza-del-Arco et al., 2021a; Plaza-del-Arco et al., 2021).

Contributions to support H2:

Contribution 4 We have identified different linguistic phenomena that might be involved in the expression of the offense.

Contribution 5 We have proposed the main methodology conducted in this doctoral thesis which follows an MTL paradigm and relies on integrating the selected linguistic phenomena in a comprehensive computational system for detecting offensive language more accurately (Plaza-del-Arco et al., 2021; Plaza-del-Arco et al., 2022).

Contributions to support H3:

Contribution 6 We have applied the proposed approach to different scenarios involved in offensive language research including sexism, hate speech, and toxicity.

Contribution 7 We have analyzed which linguistic phenomena benefit the most in each scenario through extensive experiments. We have provided a valuable discussion with the primary findings for each scenario (Plaza-del-Arco et al., 2021c; Plaza-del-Arco et al., 2021b; Plaza-del-Arco et al., 2022; Plaza-del-Arco et al., 2022)

Contribution 8 The superior performance of our proposed approach over the previous state-of-the-art approaches.

Acknowledgments

This research has been partially supported by the scholarship (FPI-PRE2019-089310) from the Ministry of Science, Innovation, and Universities, the LIVING-LANG project (RTI2018-094653-B-C21), and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

References

Aguilera-Carnerero, C. and A. H. Azeez. 2016. 'Islamonausea, not Islamophobia': The many faces of cyber hate speech. *Journal of Arab & Muslim media research*, 9(1):21–40.

Chauhan, D. S., D. S R, A. Ekbal, and P. Bhattacharyya. 2020. Sentiment and

Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online, July. Association for Computational Linguistics.

Davidson, T., D. Warmsley, M. W. Macy, and I. Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*.

Fortuna, P. and S. Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.*, 51(4), jul.

Hinduja, S. and J. W. Patchin. 2010. Bullying, cyberbullying, and suicide. *Archives of suicide research*, 14(3):206–221.

Plaza-del-Arco, F., M. Casavantes, H. Escalante, M. T. Martín-Valdivia, A. Montejo-Ráez, M. Montes-y-Gómez, H. Jarquín-Vásquez, and L. Villaseñor-Pineda. 2021. Overview of the MeOffendEs task on offensive text detection at IberLEF 2021. *Procesamiento del Lenguaje Natural*, 67.

Plaza-del-Arco, F. M., S. Halat, S. Padó, and R. Klinger. 2022. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. In *ACM SIGIR Special Interest Group on Information Retrieval TCS Research*.

Plaza-del-Arco, F. M., S. M. Jiménez-Zafra, A. Montejo-Ráez, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. 2021a. Overview of the EmoEvalEs task on emotion detection for Spanish at IberLEF 2021. *Procesamiento del Lenguaje Natural*, 67:155–161.

Plaza-del-Arco, F. M., M. D. Molina-González, U.-L. L. Alfonso, and M. V. M. Teresa. 2021b. Sinai at iberlef-2021 detoxis task: Exploring features as tasks in a multi-task learning approach to detecting toxic comments. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, volume 21, pages 580–590.

Plaza-del-Arco, F. M., M. D. Molina-González, L. López, and M. Martín-Valdivia. 2021c. Sexism identification in social networks using a multi-task learning system. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021)*.

Plaza-del-Arco, F. M., M. D. Molina-González, L. López, and M. Martín-Valdivia. 2022. Exploring the use of different linguistic phenomena for sexism identification in social networks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022)*, volume 2943, pages 491–499.

Plaza-del-Arco, F. M., M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Plaza-del-Arco, F. M., M. D. Molina-González, L. A. Ureña-López, and M.-T. Martín-Valdivia. 2022. Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. *Knowledge-Based Systems*, 258:109965.

Plaza-del-Arco, F. M., A. Montejo-Ráez, L. A. Ureña-López, and M.-T. Martín-Valdivia. 2021. OffendES: A New Corpus in Spanish for Offensive Language Research. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1096–1108, Held Online, September. INCOMA Ltd.

Plaza-del-Arco, F., A. B. Parras Portillo, P. López-Úbeda, B. Botella-Gil, and M. T. Martín-Valdivia. 2022. SHARE: A Lexicon of Harmful Expressions by Spanish Speakers. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 1307–1316, Marseille, France, June. European Language Resources Association.

Plaza-del-Arco, F., C. Strapparava, L. A. Ureña-López, and M. T. Martín-Valdivia. 2020. EmoEvent: A Multilingual Emotion Corpus based on different Events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages

1492–1498, Marseille, France, May. European Language Resources Association.

Poletto, F., V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, pages 1–47.

Tontodimamma, A., E. Nissi, A. Sarra, and L. Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126(1):157–179, January.

Wiegand, M., J. Ruppenhofer, and E. Eder. 2021. Implicitly Abusive Language – What does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online, June. Association for Computational Linguistics.

Wiegand, M., J. Ruppenhofer, and T. Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zampieri, M., S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zampieri, M., P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online), December. International Committee for Computational Linguistics.

Información General

SEPLN 2024

XL CONGRESO INTERNACIONAL DE LA SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL

24-27 de septiembre 2024
http://sepln2024.infor.uva.es/

1 *Presentación*

La XL edición del Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN) se celebrará los días 24, 25, 26 y 27 de septiembre de 2024, celebrándose el primer día una jornada de trabajo.

La ingente cantidad de información disponible en formato digital y en las distintas lenguas que hablamos hace imprescindible disponer de sistemas que permitan acceder a esa enorme biblioteca que es Internet de manera cada vez más estructurada.

En este mismo escenario, hay un interés renovado por la solución de los problemas de accesibilidad a la información y de mejora de explotación de esta en entornos multilingües. Muchas de las bases formales para abordar adecuadamente estas necesidades han sido y siguen siendo establecidas en el marco del procesamiento del lenguaje natural y de sus múltiples vertientes: extracción y recuperación de información, sistemas de búsqueda de respuestas, traducción automática, análisis automático del contenido textual, resumen automático, generación textual y reconocimiento y síntesis de voz.

2 *Objetivos*

El objetivo principal del congreso es ofrecer un foro para presentar las últimas investigaciones y desarrollos en el ámbito de trabajo del Procesamiento del Lenguaje Natural (PLN) tanto a la comunidad científica como a las empresas del sector. También se pretende mostrar las posibilidades reales de aplicación y conocer nuevos proyectos I+D en este campo.

Además, como en anteriores ediciones, se desea identificar las futuras directrices de la investigación básica y de las aplicaciones previstas por los profesionales, con el fin de contrastarlas con las necesidades reales del mercado. Finalmente, el congreso pretende ser un marco propicio para introducir a otras personas interesadas en esta área de conocimiento

3 *Áreas Temáticas*

Se anima a grupos e investigadores a enviar comunicaciones, resúmenes de proyectos o demostraciones en alguna de las áreas temáticas siguientes, entre otras:

- Desarrollo de recursos y herramientas lingüísticas.
- Análisis morfológico y sintáctico.
- Semántica, pragmática y discurso.
- Resolución de ambigüedad léxico-semántica.
- Generación de texto monolingüe y multilingüe.
- Traducción automática.
- Multimodalidad.
- Procesamiento del habla.
- Sistemas de diálogo / asistentes conversacionales.
- Indexación y recuperación de información multimedia.
- Recuperación y extracción de información monolingüe y multilingüe.
- Sistemas de búsqueda de respuestas.
- Evaluación de sistemas de PLN.
- Análisis automático de contenido textual.
- Análisis de opiniones y minería de la argumentación.
- Detección de plagio.
- Procesamiento de la negación y la especulación.
- Minería de texto en redes sociales.
- Resumen automático de texto.
- Simplificación de texto.
- Conocimiento y sentido común.
- PLN en el ámbito biomédico.

- Generación de recursos didácticos basada en PLN.
- PLN para lenguas con recursos limitados.
- Aplicaciones industriales del PLN.
- Aspectos éticos del PLN.
- Interpretabililidad y análisis de modelos para PLN.

4 *Formato del Congreso*

La duración prevista del congreso será de cuatro días, con sesiones dedicadas a la presentación de artículos, proyectos de investigación en marcha y demostraciones de aplicaciones. Además, tendrá lugar la sexta edición de IberLEF el día 24 de septiembre.

5 *Comité ejecutivo SEPLN 2024*

Presidencia del Comité Organizador

- Valentín Cardeñoso Payo (Universidad de Valladolid).
- David Escudero Mancebo (Universidad de Valladolid).

Miembros:

- Mario Corrales Astorgano (Universidad de Valladolid).
- César González Ferreras (Universidad de Valladolid).
- Eugenio Martínez Cámara (Universidad de Jaén).
- Cristina Ruiz Urbón (Universidad de Valladolid).
- Carlos E. Vivaracho (Universidad de Valladolid).

Colaboradores:

- David Fernández Martínez (Universidad de Valladolid).
- David López García (Universidad de Valladolid).

6 *Consejo Asesor*

Miembros:

- Xabier Arregi (Universidad del País Vasco, España)
- Aitziber Atutxa (Universidad del País Vasco, España)
- Miguel Ángel Alonso Pardo (Universidad de La Coruña, España)
- Manuel de Buenaga (Universidad de Alcalá, España)
- Jose Camacho Collados (Universidad de Cardiff, Reino Unido)
- Sylviane Cardey-Greenfield (Centre de recherche en linguistique et traitement automatique des langues, Francia)
- Irene Castellón (Universidad de Barcelona, España)
- Arantza Díaz de Ilarraza (Universidad del País Vasco, España)
- Antonio Ferrández (Universidad de Alicante, España)
- Koldo Gojenola (Universidad del País Vasco, España)
- José Miguel Goñi (Universidad Politécnica de Madrid, España)
- Inma Hernaez (Universidad del País Vasco, España)
- Elena Lloret (Universidad de Alicante, España)
- Ramón López-Cózar Delgado (Universidad de Granada, España)
- Bernardo Magnini Fondazione (Bruno Kessler, Italia)
- Nuno J. Mamede (Instituto de Engenharia de Sistemas e Computadores, Portugal)
- M. Teresa Martín Valdivia (Universidad de Jaén, España)
- Patricio Martínez-Barco (Universidad de Alicante, España)
- Eugenio Martínez Cámara (Universidad de Jaén, España)
- Paloma Martínez Fernández (Universidad Carlos III, España)
- Raquel Martínez Unanue (Universidad Nacional de Educación a Distancia, España)
- Ruslan Mitkov (University of Wolverhampton, Reino Unido)
- Arturo Montejo Ráez (Universidad de Jaén, España)
- Manuel Montes y Gómez (Instituto Nacional de Astrofísica, Óptica y Electrónica, México)
- Rafael Muñoz Guillena (Universidad de Alicante, España)
- Mariana Neves (German Federal Institute for Risk Assessment, Alemania)
- Lluís Padró Universidad (Politécnica de Cataluña, España)
- Manuel Palomar (Universidad de Alicante, España)
- Ferrán Pla (Universidad Politécnica de Valencia, España)

- German Rigau (Universidad del País Vasco, España)
- Álvaro Rodrigo Yuste (Universidad Nacional de Educación a Distancia, España).
- Paolo Rosso (Universidad Politécnica de Valencia, España)
- Leonel Ruiz Miyares (Centro de Lingüística Aplicada de Santiago de Cuba, Cuba)
- Horacio Saggion (Universidad Pompeu Fabra, España)
- Emilio Sanchís (Universidad Politécnica de Valencia, España)
- Encarna Segarra (Universidad Politécnica de Valencia, España)
- Thamar Solorio (University of Houston, Estados Unidos de América)
- Maite Taboada (Simon Fraser University, Canadá)
- Mariona Taulé (Universidad de Barcelona, España)
- Juan-Manuel Torres-Moreno (Laboratoire Informatique d'Avignon / Université d'Avignon, Francia)
- José Antonio Troyano Jiménez (Universidad de Sevilla, España)
- L. Alfonso Ureña López (Universidad de Jaén, España)
- Rafael Valencia García (Universidad de Murcia, España)
- René Venegas Velásques (Pontificia Universidad Católica de Valparaíso, Chile)
- Felisa Verdejo Maíllo (Universidad Nacional de Educación a Distancia, España)
- Manuel Vilares (Universidad de la Coruña, España)
- Luis Villaseñor-Pineda (Instituto Nacional de Astrofísica, Óptica y Electrónica, México)

7 *Fechas importantes*

Fechas para la presentación y aceptación de comunicaciones:

- Fecha límite para la entrega de comunicaciones: 17 de marzo de 2024.
- Notificación de aceptación: 16 de mayo de 2024.
- Fecha límite para entrega de la versión definitiva: 31 de mayo de 2024.

Información para los Autores

Formato de los Trabajos

- La longitud máxima admitida para las contribuciones será de 10 páginas DIN A4 (210 x 297 mm.), además de referencias y figuras.
- Los artículos pueden estar escritos en inglés o español. El título, resumen y palabras clave deben escribirse en ambas lenguas.
- El formato será en Word o LaTeX

Envío de los Trabajos

- El envío de los trabajos se realizará electrónicamente a través de la plataforma de envío publicada en: http://www.sepln.org/la-revista/informacion-para-autores.
- Para los trabajos con formato LaTeX se enviará el archivo PDF.
- Para los trabajos con formato Word se mandará el archivo PDF junto al DOC o RTF.
- Para más información http://www.sepln.org/index.php/la-revista/informacion-para-autores

Información Adicional

Funciones del Consejo de Redacción

Las funciones del Consejo de Redacción o Editorial de la revista SEPLN son las siguientes:

- Controlar la selección y tomar las decisiones en la publicación de los contenidos que han de conformar cada número de la revista.
- Política editorial.
- Preparación de cada número.
- Relación con los evaluadores y autores.
- Relación con el comité científico.

Funciones del Consejo Asesor

Las funciones del Consejo Asesor o Científico de la revista SEPLN son las siguientes:

- Marcar, orientar y redireccionar la política científica de la revista y las líneas de investigación a potenciar.
- Representación.
- Impulso a la difusión internacional.
- Capacidad de atracción de autores.
- Evaluación.
- Composición.
- Prestigio.
- Alta especialización.
- Internacionalidad.

Cartas al director

Sociedad Española para el Procesamiento del Lenguaje Natural
Departamento de Informática. Universidad de Jaén
Campus Las Lagunillas, Edificio A3. Despacho 127. 23071 Jaén
secretaria.sepln@ujaen.es

Más información

Para más información sobre la Sociedad Española del Procesamiento del Lenguaje Natural puede consultar la página web http://www.sepln.org.
Si desea inscribirse como socio de la Sociedad Española del Procesamiento del Lenguaje Natural puede realizarlo a través del formulario web que se encuentra en esta dirección http://www.sepln.org/sepln/inscripcion-para-nuevos-socios
Los números anteriores de la revista se encuentran disponibles en la revista electrónica: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/archive
Las funciones del Consejo de Redacción están disponibles en Internet a través de http://www.sepln.org/la-revista/consejo-de-redaccion.
Las funciones del Consejo Asesor están disponibles Internet a través de la página http://www.sepln.org/la-revista/consejo-asesor.
La inscripción como nuevo socio de la SEPLN se puede realizar a través de la página http://www.sepln.org/sepln/inscripcion-para-nuevos-socios