Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 243-256 recibido 04-07-2021 revisado 11-07-2021 aceptado 15-07-2021

NLP applied to occupational health: MEDDOPROF shared task at IberLEF 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts

PLN aplicado a salud laboral: tarea MEDDOPROF en IberLEF 2021 sobre detección, clasificación y normalización automática de profesiones y ocupaciones en textos médicos

Salvador Lima-López[1], **Eulàlia Farré-Maduell**[1], **Antonio Miranda-Escalada**[1], **Vicent Brivá-Iglesias**[2], **Martin Krallinger**[1]
[1]Barcelona Supercomputing Center, Spain
[2]Dublin City University
{salvador.limalopez, antonio.miranda, eulalia.farre, martin.krallinger}@bsc.es
vicent.brivaiglesias2@mail.dcu.ie

Resumen: Entre las características sociodemográficas de los pacientes, las ocupaciones juegan un papel fundamental tanto desde el punto de vista de la salud laboral, accidentes laborales y exposición a tóxicos y patógenos como desde el de la salud física y mental. Este artículo presenta la tarea Medical Documents Profession Recognition (MEDDOPROF), celebrada dentro de IberLEF/SEPLN 2021. La tarea se centra en el reconocimiento y detección de ocupaciones en textos médicos en castellano. MEDDOPROF propone tres retos: NER (reconocimiento de profesiones, situaciones laborales y actividades), CLASS (clasificar cada ocupación en función de su referente, como puede ser el paciente o un familiar) y NORM (normalizar menciones a las terminologías ESCO y SNOMED-CT). De un total de 40 equipos registrados, 15 han presentado un total de 94 sistemas. Los sistemas de mejor rendimiento se basan en tecnologías de aprendizaje profundo como transformers, llegando a conseguir una F-score de 0.818 en detección de ocupaciones (NER), 0.793 en clasificación de ocupaciones por su referente (CLASS) y 0.619 en normalización (NORM). Futuras iniciativas deberían tener también en cuenta aspectos multilingües y la aplicación en otros dominios como servicios sociales, recursos humanos, análisis del mercado legal y laboral o la política.
Palabras clave: tarea compartida, dominio clínico, ocupaciones, castellano.

Abstract: Among the socio-demographic patient characteristics, occupations play an important role regarding not only occupational health, work-related accidents and exposure to toxic/pathogenic agents, but also their impact on general physical and mental health. This paper presents the Medical Documents Profession Recognition (MEDDOPROF) shared task (held within IberLEF/SEPLN 2021), focused on the recognition and normalization of occupations in medical documents in Spanish. MEDDOPROF proposes three challenges: NER (recognition of professions, employment statuses and activities in text), CLASS (classifying each occupation mention to its holder, i.e. patient or family member) and NORM (normalizing mentions to their identifier in ESCO or SNOMED CT). From the total of 40 registered teams, 15 submitted a total of 94 runs for the various sub-tracks. Best-performing systems were based on deep-learning technologies (incl. transformers) and achieved 0.818 F-score in occupation detection (NER), 0.793 in classifying occupations to their referent (CLASS) and 0.619 in normalization (NORM). Future initiatives should also address multilingual aspects and application to other domains like social services, human resources, legal or job market data analytics and policy makers.
Keywords: shared task, clinical domain, occupations, Spanish.
ISSN 1135-5948. DOI 10.26342/2021-67-21

1 Introduction

Highly relevant demographic patient characteristics such as age, gender and ethnicity are often stored in structured form in electronic health records (EHRs), facilitating efficient selection and comparative statistical analysis of patient subsets. Despite the relevance of our livelihood and economic status for health and lifestyles, occupations are not systematically collected in medical documents.

Relations between disorders and occupations might respond to accidents, increased risk of contact and exposure to hazardous substances, allergens or infectious pathogens (Gambhir et al., 2011), and physical and psychological strain (Stansfeld et al., 2011). Occupational health, a medical specialty, aims to define and prevent all health issues derived from our professional activity. The characterization of the occupations of patients is extremely relevant in medicine, specifically regarding occupational transmissible and non-transmissible risks, preventive measures, health education and safety. Additionally, EHRs can contain work history information, employment statuses and activities and hobbies that while not usually structured can be very impactful (Vanotti et al., 2017).

Recently, machine learning technologies have been applied to the characterization of occupational data. Occupational data mining is a sub-field within data mining that explores different occupational variables to "build up knowledge to subsidize actions that may improve workers' quality of life" (Fernandes y Dias, 2019). It has been used, for instance, to examine the relation between workers' characteristics and workplace accidents (Gül et al., 2016) or the causes of these accidents (Cheng, Yao, and Wu, 2013). Even if some studies show that predictive models return better results when trained on narrative texts rather than tabular data (Yedla, Kakhki, and Jannesari, 2020), there are not many openly available resources that allow a thorough analysis of this type of data.

Automatic extraction of mentions of occupations and employment status from unstructured text is a Named Entity Recognition (NER) task, where predefined types of concepts are detected directly from documents. Prior NER benchmark efforts traditionally focused on general entities like person names, organizations and locations, while more specific entities were left for specialized systems. Although the Sixth Message Understanding Conference (muc, 1995) already included the detection of management posts (a very specific type of occupation), recent NER strategies have only marginally addressed this entity type when considering more specialized domains, text genres or non-English content. Nevertheless, there are some exceptions, such as the German NoSta-D dataset (Benikova, Biemann, and Reznicek, 2014), which considered professions as a type of organizations.

In clinical text processing, the detection of profession mentions was partially addressed by de-identification shared task efforts. Several shared tasks regarded occupations as a type of Personal Health Information item that had to be detected, hidden from plain text or pseudo-anonymized. In English, the 2014 i2b2/UTHealth (Stubbs y Uzuner, 2015) and the 2016 CEGS N-GRID (Stubbs, Filannino, and Uzuner, 2017) shared tasks both included occupations in their datasets. For Spanish, the MEDDOCAN track of IberEval 2019 (Marimon et al., 2019) also included occupations. However, occupations were among the most difficult types of Personal Health Information when examining the results obtained by automatic systems (Stubbs y Uzuner, 2015). Outside de-identification, recently the Industrial and Professional Occupations Dataset (IPOD) (Liu et al., 2020) was released, which includes a corpus of 475,085 job titles in English crawled from LinkedIn and a gazetteer.

The limitation of these resources is that they only consider occupations when job titles are directly mentioned. However, natural languages use many expressions to describe occupations. In narrative texts, including clinical documents, mentions of occupations include a description of its mission, the materials used by the worker, a reference to the workplace and other. In order to detect all these occupational nuances, more exhaustive resources that reflect how speakers actually talk about occupations are needed.

An interesting exception to this problem is SkillNER (Fareri et al., 2021), a named entity recognition system trained to detect workers' skills and competences. Terminological resources like ESCO (European Skills, Competences, Qualifications and Occupations) (European Commission, 2013) – a multilingual classification based upon the International Standard Classification of Occupations

(ISCO) (International Labour Organization, 2021) – also give some more insight into the occupation ecosystem. ESCO "identifies and categorises skills, competences, qualifications and occupations relevant for the EU labour market and education and training", as well as their relations, by providing each with a unique identifier. Clinical terminologies like SNOMED-CT also include a myriad of occupations, highlighting the importance of these variables in medicine. To the best of our knowledge, there are no other available resources that describe occupation mentions in depth.

Characteristics that profession and employment status recognition systems need to address in clinical texts include: (1) heterogeneous types of profession-relevant descriptions, (2) classification of whether the detected expressions refer to the patient, a family member, or a healthcare professional and (3) mapping the extracted mentions to controlled vocabularies or normative terminologies to enable semantic interoperability. The complexity of this task lies in the lack of precursors, the wide range of expressions referring to occupations, and the descriptive and heterogeneous nature of medical documents. Systems able to address these issues are also crucial for social workers, the industry, policy makers and Human Resources departments.

In order to find solutions and following the success of previous Information Extraction shared tasks in Spanish like CANTEMIST (Miranda-Escalada, Farré, and Krallinger, 2020) and MEDDOCAN (Marimon et al., 2019), we have organized the MEDDOPROF shared task (MEDical DOcuments PROFession Recognition). MEDDOPROF aims to foster the creation of resources and occupations detection systems in the field of occupational data mining. It is the second shared task on occupational text mining in Spanish, after the social media-focused ProfNER (Miranda-Escalada et al., 2021).

2 Task Description

2.1 Shared Task Goal

MEDDOPROF focused on the detection and normalization of occupation mentions in clinical case reports written in Spanish. To facilitate extrapolation to EHRs, the MEDDOPROF clinical cases were compared using text similarity strategies to different clinical records including discharge summaries. Additionally, to maximize the practical impact of this task, participating teams were asked (1) to detect different types of occupations, namely professions, employment status and non-paid activities and (2) to classify them according to who they refer to in the text (patient, family member, health professional, someone else). We also asked them to normalize the detected mentions to one of two highly used multilingual controlled vocabularies: the European Skills, Competences, Qualifications and Occupations classification (ESCO) and SNOMED-CT. Figure 1 gives a general overview of the task.

2.2 Sub-tracks

The shared task was divided into three subtracks, each of them associated to different practically relevant use cases:

- *MEDDOPROF-NER track.* Participants were asked to find exact mentions of occupations in the text and label them according to the type of occupation: profession, employment status or activity.
- *MEDDOPROF-CLASS track.* This subtrack required finding mentions of occupations in the text as well as determining the person(s) referred to (patient, family member, health professional, someone else).
- *MEDDOPROF-NORM track.* Given a list of valid codes that includes all of ESCO and a selection of SNOMED-CT terms, participants were asked to automatically normalize the detected entity mentions to their corresponding concept identifier. This sub-track is highly relevant to enable semantic interoperability, data integration and practical exploitation of NER text mining systems.

2.3 Shared Task Setting

The MEDDOPROF shared task had two distinct phases:

Training phase. The training set was released in April 2021 and participants had almost two months to build their systems.

Evaluation phase. In June 2021, the test set was released without annotations. Participants had around two weeks to make their predictions and submit them. Each team was allowed up to five runs per sub-track.

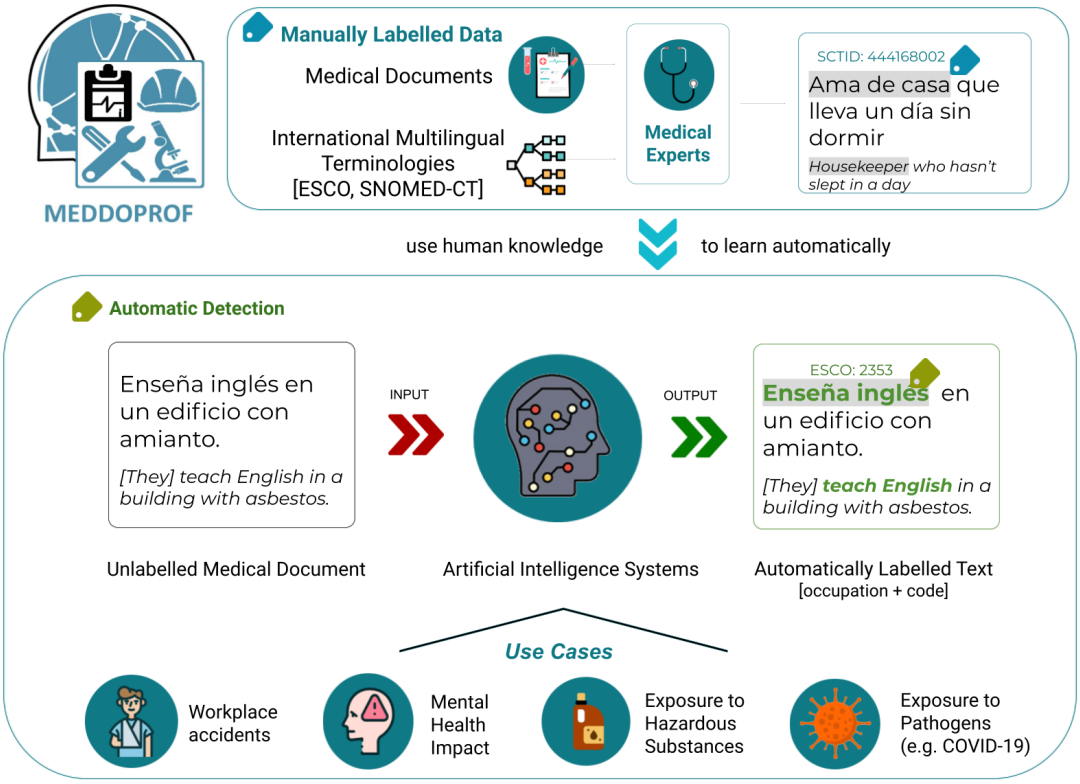

Figure 1: MEDDOPROF shared task overview.

2.4 Evaluation: Metrics and Baseline

All three sub-tracks are evaluated using micro averaged precision, recall and F-1 score. A strict evaluation setting is followed, i.e. only exact character offset matches between predictions and the manually labeled Gold Standard are considered correct. In sub-tracks MEDDOPROF-NER and MEDDOPROF-CLASS, this means that both the text span and label had to be correct. In sub-track MEDDOPROF-NORM, the predicted codes had to correspond to the manually assigned concept codes from the Gold Standard corpus; thus, parent/children codes are not considered correct predictions.

As a baseline system, we implemented a Levenshtein lexical lookup system with a sliding window. The system uses the annotations from the training set and scans the input text to find new matches. For sub-track 3, the predictions generated by the lexical lookup were compared to the training data to predict codes. The results are shown in Table 1 and the code is available on GitHub[1].

Sub-track	Precision	Recall	F-Score
NER	0.465	0.508	0.486
CLASS	0.391	0.377	0.384
NORM	0.502	0.533	0.517

Table 1: Baseline results.

3 Corpus and Resources

3.1 MEDDOPROF Gold Standard

The Gold Standard corpus for MEDDOPROF is a collection of 1844 clinical cases, reports of individual patients published in medical journals. Unlike Electronic Health Records (EHRs), which might include a lot of personal information of patients and cannot usually be shared due to privacy issues, clinical cases are already anonymized. The MEDDOPROF corpus documents were manually selected to account for a wide range of occupations and clinical specialities, including occupational medicine, primary care, COVID-19 and psychiatry. The MEDDOPROF corpus was split into training (around 80%) and test (around 20%) sets. The complete Gold Standard corpus is available at Zenodo[2].

The corpus' entities are distributed along two different annotation axes, one related to

[1]https://github.com/TeMU-BSC/meddoprof-baseline

[2]https://doi.org/10.5281/zenodo.5070541

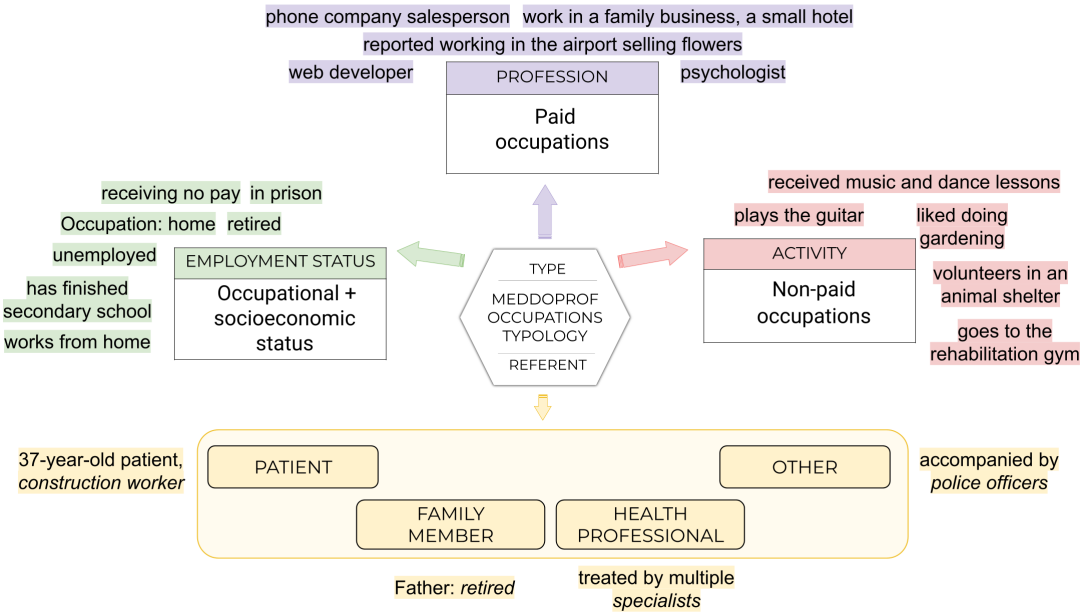

Figure 2: Overview of the categories in the MEDDOPROF Gold Standard corpus.

occupations and another related to the actual person it refers to. Three different types of occupations were considered: a) PROFESION (*profession*; paid jobs), b) SITUACION_LABORAL (*employment status*; occupational and socioeconomic statuses), and c) ACTIVIDAD (*activities*; non-paid tasks such as hobbies). Each mention was subsequently classified according to its referent in the text as: a) PACIENTE (*patient*), b) FAMILIAR (*family member*), c) SANITARIO (*health worker*) and d) OTRO (*other*). This classification is visually explained, along with some examples, in Figure 2.

3.1.1 Annotation and Normalization Process

The MEDDOPROF Gold Standard corpus was manually annotated by a linguist in collaboration with a clinician. Based on our experience with previous tasks (e.g. ProfNER (Miranda-Escalada et al., 2021) and MEDDOCAN (Marimon et al., 2019)), we conducted an iterative annotation quality control and refinement process. The first step involves the creation and definition of specific guidelines that aim to include and constrain how occupations are expressed in clinical documents. These guidelines were refined through various annotation rounds by two different annotators. The iterations continued until a high Inter-Annotator Agreement (pairwise) was reached. Finally, around 15% of the documents were sistematically cross-checked by an internal annotator for consistency, reaching an IAA of 0.9. The annotation tool brat was used (Stenetorp et al., 2012).

Regarding MEDDOPROF corpus normalisation, we first tried to find an ESCO code that reflected closely the profession/activity mentioned in the clinical reports. When we did not find an ESCO code, as was the case for most employment statuses, we turned to SNOMED-CT codes, which proved to be quite comprehensive and useful regarding employment status situations (retired, unemployed, homeless, etc). A general overview of the normalisation process and some common codes are described in the guidelines.

The final version of the MEDDOPROF guidelines is 33 pages long and includes over 70 rules and exceptions. They are freely available at Zenodo[3].

3.1.2 Corpus Format

For the first two sub-tracks (NER and CLASS), the MEDDOPROF corpus' clinical cases are provided as UTF-8 text files with the annotations as separate files in brat standoff format (.ann files). For each clinical case, there is an associated .ann file with the same name in each sub-track.

For the normalization sub-track (NORM), mappings are distributed in a tab-separated file that includes four columns: filename, mention text, span and code. For reference,

[3]https://doi.org/10.5281/zenodo.4720833

	Documents	Annotations	Unique Codes	Sentences	Tokens
Train	1,500	3,658	297	49,114	1,075,655
Test	344	1,085	167	9,513	215,531
Total	1,844	4,657	346	58,627	1,291,186

Table 2: General distribution of the MEDDOPROF corpus.

a list containing all valid codes from the used vocabularies (ESCO and SNOMED-CT) was also shared.

3.1.3 Corpus Statistics

The MEDDOPROF corpus is made up of 1,844 documents, with a total of 58,627 sentences and 1,291,186 tokens. A complete corpus overview is provided in Table 2, while Table 3 breaks down the label distribution of the corpus.

In total, there are 4,743 annotations, of which 2,058 are unique mention strings. On average, each document had between two and three annotations. Each mention was manually mapped to either ESCO or SNOMED-CT, resulting in a total of 3,191 annotations mapped to ESCO (297 unique codes) and 1,552 mapped to SNOMED-CT (49 unique codes). More details and the challenges of the annotation and normalization process are discussed in Section 5.

3.2 MEDDOPROF additional resources

Complementary Entities Dataset. In order to connect the occupational and clinical aspects of the text, a version of the 1500 documents of the training set that includes automatic annotations for clinical and linguistic variables was released. This dataset was tagged using an adapted and retrained version of the system previously used for the PharmacoNER Tagger (Amengol-Estapé et al., 2019) and in-house manually labelled corpora. It includes the following entities: symptoms, diseases, procedures, drugs, species, negation trigger and scope and uncertainty trigger and scope. Participants were free to implement this extra knowledge in their systems. Table 4 gives a general view of the entities in this additional dataset, while Figure 3 provides an example an annotated clinical case from this corpus.

The MEDDOPROF complementary entities dataset can be downloaded together with the training set at Zenodo [4].

Occupations Gazetteer. A gazetteer of occupations in Spanish was also released as a tab-separated file. This resource includes over 25,000 terms and was constructed by extracting information from terminological resources from multiple terminologies (DeCS, ESCO, SnomedCT and WordNet) and occupations detected by Stanford CoreNLP in a large collection of social media Spanish profiles. The gazetteer can be found in Zenodo[5].

4 Results

MEDDOPROF contains three sub-tracks: MEDDOPROF-NER, MEDDOPROF-CLASS and MEDDOPROF-NORM. Only the first two tracks are independent, as NORM requires the output of at least one of the other tracks. Participants could choose to submit results for one, two or all three sub-tracks. Up to five submissions were allowed for each sub-track. This section summarizes the task's participation, results and methodologies used.

4.1 Participation overview

A total of 40 teams registered for the task, of which 15 submitted their predictions. All 15 teams participated in the MEDDOPROF-NER sub-track, 11 in the MEDDOPROF-CLASS sub-track and 8 in the MEDDOPROF-NORM sub-track. Due to the fact that up to 5 systems could be submitted, the final number of prediction runs is high, a total of 94: the NER sub-track received 39 runs, CLASS received 29 runs and NORM received 26 runs.

MEDDOPROF attracted participants from very diverse backgrounds. Ten teams were from academia, three were from industry, one is a collaboration between academia an industry and one participant is a freelance. Even though most teams were from Spain, there were also MEDDOPROF participants from France, Germany, India, Mexico and the United Kingdom. Table 5

[4] https://doi.org/10.5281/zenodo.4775741

[5] https://doi.org/10.5281/zenodo.4524659

	Patient	Family	Health Prof.	Other	Total
Profession	1,158	134	1,525	410	3,227
Empl. Status	1,047	119	0	203	1,369
Activity	122	7	0	18	147
Total	2,327	260	1,525	631	4,743

Table 3: Label distribution of the MEDDOPROF corpus.

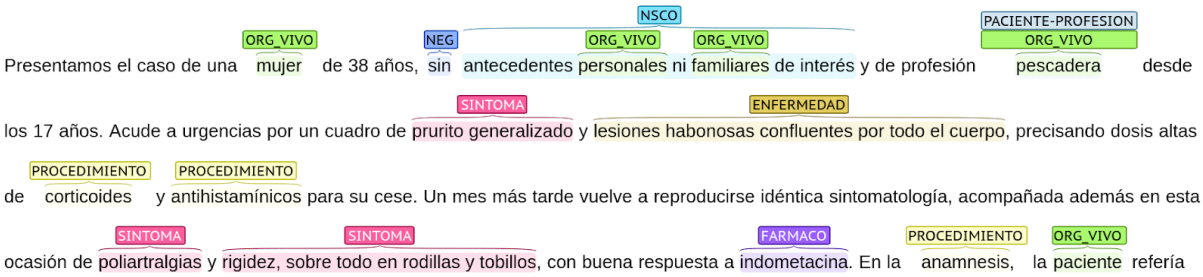

Figure 3: Example of a clinical case from the MEDDOPROF test set annotated with complementary clinical and linguistic entities.

Entity	Frequency
Síntoma (*Symptom*)	17,009
Enfermedad (*Disease*)	20,889
Procedimiento (*Procedures*)	15,033
Fármaco (*Drug*)	7,960
Org. Vivo (*Living Being*)	23,150
NEG (*Neg. Trigger*)	18,043
NSCO (*Neg. Scope*)	15,968
UNC (*Unc. Trigger*)	2,912
USCO (*Unc. Scope*)	2,765

Table 4: Distribution of the Complementary Entities Dataset.

shows an overview of the MEDDOPROF participant teams.

4.2 System results

Table 6 shows the best results obtained by each team. The top results for each sub-track were:

- *MEDDOPROF-NER*. The best performing system for this sub-track was implemented by the team NLNDE, with an F1-score of 0.818. Their precision was 0.855 and their recall was 0.783. They used a combination of pretrained transformers with Masked Language Modelling (MLM) and CRFs, as well as strategic data splits. The second best system was submitted by the MUCIC team, who obtained a 0.8 F1-score and employed a model based on Flair with BERT embeddings.
- *MEDDOPROF-CLASS*. NLNDE obtained the best F1-score (0.793), along with the best precision (0.83) and recall (0.759). MUCIC's submission was also the second best system, with an F1-score of 0.764. Both teams employed the same system described for the NER sub-track.
- *MEDDOPROF-NORM*. The TALP team obtained the best results in this task using a pre-trained multilingual DistilBERT (Sanh et al., 2019) plus a BiLSTM. They achieved an F1-score of 0.619, a precision of 0.675 and a recall of 0.572. Fadi obtained the second best result with an F1-score of 0.603 with a BERT transformer.

For a complete overview of the task's results, please refer to the Annex at the end.

4.3 Methodologies

Over the past few years, some of the biggest advancements in NLP have been achieved by large neural language models and their transformer architecture. This trend has been observed in MEDDOPROF, where the most common architecture used by participants are transformers-based language models, mainly BERT (Devlin et al., 2019) or its Spanish version BETO (Cañete et al., 2020). In all three sub-tracks, almost every team used them either directly – fine-tuning the task data, or in the form of embeddings.

Conditional Random Fields were also utilized by a few teams, sometimes as a complete system (gbali team) and sometimes as

Team Name	Affiliation	Country	A/I	Tasks	Ref.
EdIE-KnowLab	University of Edinburgh	UK	A	NE,CL,NO	(Suárez-Paniagua y Casey, 2021)
Fadi	Universitat Rovira i Virgili	Spain	A	NE,CL,NO	-
Galiza	Universitat Oberta de Catalunya	Spain	A	NE,NO	-
gbali	Independent	France	-	NE,CL	-
HULAT-UC3M	Universidad Carlos III de Madrid	Spain	A	NE	-
ICC	Instituto de Ingeniería del Conocimiento	Spain	A	NE,CL,NO	-
IITKGP	Karaghpur Indian Institute of Technology	India	A	NE	(Harkawat y Vaidhya, 2021)
KaushikAcharya	Philips India Limited	India	I	NE,NO	(Acharya, 2021)
MUCIC	Center for Computing Research, National Polytechnic Institute	Mexico	A	NE,CL	(Balouchzahi, Sidorov, and Shashirekha, 2021)
NLNDE	Bosch Center for Artificial Intelligence	Germany	I	NE,CL	(Lange, Adel, and Strötgen, 2021)
SINAI	Universidad de Jaén	Spain	A	NE,CL,NO	(Mesa-Murgado et al., 2021)
SMR-NLP	Siemens AG / Ludwig Maximilian University of Munich	Germany	I	NE,CL	-
TALP	Universitat Politècnica de Catalunya	Spain	A	NE,CL,NO	(Medina Herrera y Turmo Borràs, 2021)
URJC-UNED Team	Universidad Rey Juan Carlos / Universidad Nacional de Educación a Distancia	Spain	A	NE,CL	-
Vicomtech NLP-Team	Vicomtech Foundation	Spain	I	NE,CL,NO	(Zotova, García-Pablos, and Cuadros, 2021)

Table 5: MEDDOPROF team overview. A/I stands for academic or industry institution. In the Tasks column, NE stands for MEDDOPROF-NER, CL for MEDDOPROF-CLASS and NO for MEDDOPROF-NORM.

	NER			CLASS			NORM		
Team Name	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
EdIE-KnowLab	0.585	0.712	0.643	0.604	0.604	0.604	0.165	0.193	0.178
Fadi	0.802	0.678	0.735	0.761	0.644	0.698	0.682	0.541	0.603
Galiza	0.731	0.597	0.657	-	-	-	0.72	0.482	0.577
gbali	0.786	0.586	0.671	0.726	0.538	0.618	-	-	-
HULAT-UC3M	0.412	0.53	0.464	-	-	-	-	-	-
ICC	0.741	0.435	0.549	0.662	0.377	0.48	0.567	0.388	0.461
IITKGP	0.654	0.5	0.567	-	-	-	-	-	-
KaushikAcharya	0.807	0.524	0.635	-	-	-	**0.72**	0.467	0.566
MUCIC	0.813	**0.788**	0.8	0.77	0.75	0.764	-	-	-
NLNDE	**0.855**	0.783	**0.818**	**0.83**	**0.759**	**0.793**	-	-	-
SINAI	0.821	0.74	0.778	0.775	0.69	0.73	0.593	0.541	0.566
SMR-NLP	0.854	0.751	0.799	0.802	0.699	0.747	-	-	-
TALP	0.761	0.465	0.698	0.694	0.588	0.637	0.675	**0.572**	**0.619**
URJC-UNED Team	0.765	0.706	0.734	0.71	0.664	0.686	-	-	-
Vicomtech NLP-team	0.758	0.739	0.748	0.71	0.691	0.701	0.488	0. 474	0.481
Baseline	*0.465*	*0.508*	*0.486*	*0.391*	*0.377*	*0.384*	*0.502*	*0.533*	*0.517*

Table 6: Best result per team. The best result in each sub-track is presented in bold letters, the second best is underlined. A dash indicates that the team did not participate in that sub-track. For reference, P indicates Precision, R indicates Recall and F1 indicates F1-score.

a final classification layer of another architecture (NLNDE team). One team (SMR-NLP) tried a Bidirectional Recurrent Neural Network with Long Short Team Memory (BiLSTM). Teams like NLNDE also experimented with variables like domain-specific models and splitting the data into strategic partitions with great success.

There were also some entries that used alternative methods to neural systems. For instance, one of the submissions of the Galiza team was a non-neural lookup system based on rules. Another team, HULAT-UC3M, proposed a system that used MeaningCloud's topic extraction API (mea, 2021) to build a dictionary-based search engine.

It is also noteworthy that most participants used well-known NLP libraries such as

spaCy (Honnibal et al., 2020), HuggingFace (Wolf et al., 2020) or Flair (Akbik et al., 2019), either to build the entire system or at some point in their pipeline.

As for the choice of systems for each sub-track, many teams re-used their models for all three challenges, either by training the model to output multiple labels or via transfer learning. Even then, results for the NER sub-track are overall higher than for the CLASS sub-track, although the data used for both is the same. This suggests that the CLASS sub-track is somewhat harder, something that is to be expected given that matching an occupation to its holder requires a good understanding of the sentence's semantics and syntactic structure. For the normalization sub-track, many teams also employed the same systems as for the other sub-tracks. Some teams, like the Vicomtech NLP-team, opted to build more specialized systems that use semantic search.

4.4 Error analysis

This section goes through some of the most frequent errors observed when comparing the Gold Standard annotations and the predictions of the participants' systems.

Ambiguous words. In Spanish, many occupation terms are polysemous or homonyms and can be used as nouns and adjectives. For instance, the words "médico" and "clínico" both mean *doctor* as a noun and *clinical* as an adjective. This distinction seems to be somewhat complex for current systems to understand and it is a common source of false positives. This type of error happens a lot with many frequent words: "profesional" (*professional*), "ejecutivo" (*executive*), "ganadero" (*stock breeder*), "informático" (*computer scientist*), "matemático" (*mathematician*), ...

Something similar happens with the word "compañeros" (*partner*, *colleague*). It can refer to "compañeros de trabajo" (*coworker*), "compañeros de colegio" (*classmates*), "compañeros de piso" (*flatmates*) and so on. These mentions are highly context-dependent, and in the Gold Standard they were annotated only when they related to occupations. However, some systems labeled all instances of the word regardless of the context.

An interesting point is that some systems that use morphological features predicted unrelated words that sound like common occupations. For instance, one system decided that "platanero" (*banana tree*, which sounds like "camarero", *waiter*) or "desfibrador" (*defibrator*, which sounds like "leñador", *woodcutter*) were professions.

Ambiguous mentions. In the Gold Standard, there are some frequent mentions that appear in more than one category. The two main ones are:

- "Trabajador" (*worker*). On its own, without any specification, this word is considered an employment status. However, whenever it is accompanied by some type of job description (workplace, specialty), it is considered a profession (as in "trabajador del sector metalúrgico", *metalworker*). Many systems falsely labelled "trabajador" on its own as a profession, failing to consider the context and, thus, not including the job description in the prediction. The opposite is also true: some predictions included an occupation together with an adjectival or prepositional phrase that is not part of the occupation.

- "Cuidador" (*caregiver*). Most caregivers are either a paid profession or a role taken up by a family member, in which case it was annotated as employment status. This distinction requires a good understanding of the sentence's context and thus proved to be difficult.

This situation was also replicated in the case of activities that can also be paid occupations such as "deportista" (*sports player*) or "músico" (*musician*).

Scope. One of the main challenges of NER systems is correct boundary detection (Li et al., 2020). Due to the variety in the corpus' annotations, predictions being too short or too long are often a source of errors. On the one hand, some common examples of a prediction being too short are:

Affixes. Some prefixes like "ex-" were included as part of the annotated mentions as they provide important information. However, when they were separated by a space rather than joined together with the word they accompany (e.g. "ex trabajadora de fábrica textil" *former textile factory worker*), multiple systems did not include the prefix.

Occupational granularity. Some occupational terms are understood on their own or include information about their specialty, workplace, ... (e.g. "ingeniero" *engineer* vs. "ingeniero de caminos" *civil engineer*). Some predictions included only the standalone occupation and failed to include the specifics.

Coordinated phrases. Some occupations may reference more than one specialty, as in "profesional de la estética y la belleza" (*aesthetics and beauty professional*). There were systems who stopped at the first specialty and did not include the coordinated part.

On the other hand, some predictions were too long for reasons such as:

Syntactically similar structures. Given the variety of occupational references in the corpus, a good system needs to understand the semantics of the sentence to distinguish relevant from irrelevant information in syntactically similar structures. This was not always the case, as some predictions include irrelevant information due to their syntactic distribution. For instance, in the sentence "[...] clasifica al trabajador en tres categorías: apto / no apto / apto en determinadas condiciones" (*[...] classify the worker in three categories: apt / not apt / apt under certain conditions*), one system predicted "trabajador en tres categorías" as if "tres categorías" was a specialty. Another example is "profesor en Costa de Marfil" (*teacher in Ivory Coast*), where "Costa de Marfil" was predicted even though it is a geographical location.

Separate mentions joined together. Sometimes, a profession may appear together with an employment status or activity, as in "agricultor jubilado" (*retired farmer*). These cases were annotated separately, but were at times predicted as a single entity.

In for the normalization task the main source of errors was **code granularity**. Multiple systems predicted a parent/child concept node of the GS code, but only exact codes were considered correct. This phenomenon may be influenced by the similarity between some of ESCO's concepts.

5 Discussion

We present the MEDDOPROF shared task results and resources on the automatic detection and normalization of occupations from medical documents written in Spanish. To the best of our knowledge, it is the first attempt at characterizing occupations in clinical documents. Most NLP research has considered only job titles as occupations, neglecting the considerable variability of language expressions used to refer to occupational information in written texts. Robust detection of occupation information in text requires a thorough characterization that goes beyond what gazetteer or dictionary-based resources can handle.

Although MEDDOPROF deals with documents in Spanish, it is clear that the release of the MEDDOPROF annotation guidelines and the use of multilingual terminologies for normalization can serve as base for similar efforts in other languages and other application domains.

The following resources have been released as part of MEDDOPROF: a normalized Gold Standard corpus that characterizes occupational language in medical documents in Spanish[6], 33 pages long annotation guidelines that describe how to annotate this phenomenon[7], a version of the training set (named MEDDOPROF Complementary Entities) that includes automatic predictions of clinical and linguistic variables[8] and a gazetteer of over 25,000 occupational terms[9]. High quality annotation guidelines are key to allow extending the corpus beyond the current size and to be able to interpret and understand the manual and automatic annotations preventing a *Black box corpus scenario* often encountered in NLP research.

A total of 15 participants submitted at least one system, achieving 0.818 F-score in occupation detection (MEDDOPROF-NER), 0.793 in classifying occupations to their holder (MEDDOPROF-CLASS) and 0.619 in normalization (MEDDOPROF-NORM). Given the somewhat small dataset size and the task's complexity (including highly ambiguous and context-dependent mentions), the results are promising. In terms of each specific sub-track, the results for the NER track are the highest overall. The CLASS sub-track's results are somewhat lower despite using the same data, which indicates that it is more complex and harder to learn.

We anticipate that the systems resulting from the MEDDOPROF task will highlight the importance of socio-demographic entities

[6] https://doi.org/10.5281/zenodo.5070540
[7] https://doi.org/10.5281/zenodo.4694675
[8] https://doi.org/10.5281/zenodo.4694768
[9] https://doi.org/10.5281/zenodo.4524658

like occupations, especially in the clinical domain, and help in their processing. Beyond the healthcare application domain, we expect that systems similar to those used for the MEDDOPROF track may be adapted and applied to heterogeneous fields such as social care, human resources, legal NLP and even gender studies. European research projects like the Exposome Project for Health and Occupational Research (EPHOR) could also benefit from this type of resources and systems. Finally, other NLP tasks such as anonymization might also benefit from having more exhaustively annotated data.

Despite the use of two comprehensive controlled vocabularies (ESCO and SNOMED CT), entity normalization was challenging, as it had to manage concept code granularity and concept ambiguity. Some of the entries in ESCO were highly similar, hindering the selection of the most appropriate code. For example, codes 01, 011 and 0110 are all named "commissioned armed forces officers". The difference between them is given as a short text description, which is difficult to exploit for automatic tools. In case of mentions that could potentially fit several candidate concepts (e.g. "militar" *soldier*), we opted for mapping to the least granular code.

A number of employment situations, activities and some illegal occupations were not covered by ESCO. For such entity types, we used SNOMED-CT as a target vocabulary. Nonetheless, it was cumbersome to normalize some important activities of social and medical significance, including volunteering in charities, and many physical activities practised regularly (which we had to label with the general code SCTID: 228447005 "Physically active (finding)"). Similarly, there was no appropriate code for people who regularly played instruments or went to dance lessons, which were simply ascribed as SCTID:300758009 ("Does engage in a hobby (finding)"). Notably, while we do not dispute the therapeutic effects of these activities, the most approximate codes for regular practice of yoga (SCTID:229224000, "Participation in yoga (regime/therapy)") and pilates (SCTID:404928000, "Pilates exercise (regime/therapy)") necessarily attached a medical significance to these mentions.

Current trends of preventive medicine and healthy aging include the practice of many activities like going to the gym, singing in choirs, going for group walks and volunteering as part of healthy lifestyles. Primary care centres assess the neighbourhood assets and general practitioners use social prescribing to improve the health, well-being and quality of life of patients. We believe that these activities should be properly annotated and normalised in medical records to obtain a comprehensive understanding of the health of the patient, to further personalised medicine and to investigate the most beneficial activities for different population groups.

Usually, clinical named entity recognition requires a very strict evaluation setting due to the delicate nature of the clinical entities. Looking at the participant systems' False Positives, there are many predictions that are almost correct. Sometimes, the difference between the Gold Standard and the predictions is simply a comma. Some other times, the predictions are approximately correct: they include the GS annotation but also a noun modifier or prepositional phrase that was not captured in the GS because it was not relevant.

There are some of these approximate predictions which might look perfectly okay to a human even if they do not exactly match the GS. Because of this, in future tasks we would like to look into integrating different evaluation methods such as partial matching. This has been done in previous shared tasks, for instance Task 9 ("Extraction of Drug-Drug Interactions from BioMedical Texts") in SemEval 2013 (Segura-Bedmar, Martínez, and Herrero-Zazo, 2013).

As for the normalization's evaluation, the results show that there were systems that had problems choosing the right granularity within a family of codes. We initially considered using a ranked evaluation, but given the uneven difference in granularity of some ESCO codes and the use of more than one terminology, we ultimately decided against it and also used a strict setting. Following the exploratory semantic similarity evaluation criteria already used for the MESINESP2 shared task (Gasco et al., 2021), such metrics could provide alternative approaches to exploit better the different levels of granularity and structure of controlled vocabularies during the evaluation process.

Acknowledgements

MEDDOPROF was promoted through the collaboration between the Spanish Plan for the Advancement of Language Technology (Plan TL) and the BSC. We also want to acknowledge the 2020 Proyectos de I+D+i - RTI Tipo A (DESCIFRANDO EL PAPEL DE LAS PROFESIONES EN LA SALUD DE LOS PACIENTES A TRAVES DE LA MINERIA DE TEXTOS (PID2020-119266RA-I00)) for support. Finally, we would like to thank: the MEDDOPROF scientific committee, in special Michelle Turner (ISGlobal) and Francisco Javier Sanz Valero (Escuela Nacional de Medicina del Trabajo, Instituto de Salud Carlos III), as well as Marvin Agüero-Torales, Luis Gascó Sánchez and everyone who participated in the task.

Bibliography

1995. Appendix F: Information extraction task: Scenario on management succession(vl.l). In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

2021. MeaningCloud.

Acharya, K. 2021. Occupation Recognition and Normalization in Clinical Notes. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings*.

Akbik, A., T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Amengol-Estapé, J., F. Soares, M. Marimon, and M. Krallinger. 2019. Pharmaconer tagger: a deep learning-based tool for automatically finding chemicals and drugs in spanish medical texts. *Genomics Informatics*, 17:e15, 06.

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 *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings*.

Benikova, D., C. Biemann, and M. Reznicek. 2014. NoSta-D named entity annotation for German: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland, Mayo. European Language Resources Association (ELRA).

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Cheng, C.-W., H.-Q. Yao, and T.-C. Wu. 2013. Applying data mining techniques to analyze the causes of major occupational accidents in the petrochemical industry. *Journal of Loss Prevention in the Process Industries*, 26(6):1269–1278.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota, Junio. Association for Computational Linguistics.

European Commission. 2013. European skills/competences, qualifications and occupations.

Fareri, S., N. Melluso, F. Chiarello, and G. Fantoni. 2021. Skillner: Mining and mapping soft skills from any text. *CoRR*, abs/2101.11431.

Fernandes, F. and A. Dias. 2019. Perspectivas do uso de mineração de dados e aprendizado de máquina em saúde e segurança no trabalho. *Revista Brasileira de Saúde Ocupacional*, 44, 01.

Gambhir, R., G. Singh, S. Sharma, R. Brar, and H. Kakar. 2011. Occupational health hazards in current dental profession-a review. *The Open Occupational Health and Safety Journal*, 3, 12.

Gasco, L., A. Nentidis, A. Krithara, D. Estrada-Zavala, , R.-T. Murasaki, E. Primo-Peña, C. Bojo-Canales, G. Paliouras, and M. Krallinger. 2021. Overview of BioASQ 2021-MESINESP track. Evaluation of advance hierarchical

classification techniques for scientific literature, patents and clinical trials.

Gül, M., A. Guneri, Y. Fatih, and O. Çelebi. 2016. Analysis of the relation between the characteristics of workers and occupational accidents using data mining. *The Turkish Journal of Occupational / Environmental Medicine and Safety*, 1:102–118, 04.

Harkawat, J. and T. Vaidhya. 2021. Analysis of the Spanish Pre-Train Language Model for HealthCare Name Entity Recognition. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings*.

Honnibal, M., I. Montani, S. Van Landeghem, and A. Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

International Labour Organization. 2021. International standard classification of occupations (ISCO).

Lange, L., H. Adel, and J. Strötgen. 2021. NLNDE at MEDDOPROF: Boosting Transformers in a Low-Resource Setting. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings*.

Li, J., A. Sun, J. Han, and C. Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, J., Y. C. Ng, K. L. Wood, and K. H. Lim. 2020. Ipod: A large-scale industrial and professional occupation dataset. In *Proceedings of the 2020 ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW'20)*, pages 323–328.

Marimon, M., A. Gonzalez-Agirre, A. Intxaurrondo, H. Rodriguez, J. L. Martin, M. Villegas, and M. Krallinger. 2019. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *IberLEF@SEPLN*, pages 618–638.

Medina Herrera, S. and J. Turmo Borràs. 2021. Everything Transformers: Recognition, Classification and Normalisation of Professions and Family Relations. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings*.

Mesa-Murgado, J.-A., P. López-Úbeda, M.-C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Ureña-López. 2021. BERT Representations to Identify Professions and Employment Statuses in Health data. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings*.

Miranda-Escalada, A., E. Farré, and M. Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*.

Miranda-Escalada, A., E. Farré-Maduell, S. Lima-López, L. Gascó, V. Briva-Iglesias, M. Agüero-Torales, and M. Krallinger. 2021. The profner shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 13–20.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Segura-Bedmar, I., P. Martínez, and M. Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA, Junio. Association for Computational Linguistics.

Stansfeld, S., F. Rasul, J. Head, and N. Singleton. 2011. Occupation and mental health in a national uk survey. *Social psychiatry and psychiatric epidemiology*, 46:101–10, 02.

Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Stubbs, A., M. Filannino, and O. Uzuner. 2017. De-identification of psychiatric intake records: Overview of 2016 cegs n-grid shared tasks track 1. *Journal of Biomedical Informatics*, 75, 06.

Stubbs, A. and O. Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58S, 08.

Suárez-Paniagua, V. and A. Casey. 2021. BERT and Approximate String Matching for Automatic Recognition and Normalization of Professions in Spanish Medical Documents. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings.*

Vanotti, S., M. Eizaguirre, C. Yastremiz, A. Marinangeli, R. Alonso, B. Silva, A. Iorio, F. Cáceres, and O. Garcea. 2017. Estudio del estatus laboral y el nivel socioeconómico en personas con esclerosis múltiple en 2 centros de buenos aires. *Neurología Argentina*, 10, 09.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, 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. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Octubre. Association for Computational Linguistics.

Yedla, A., F. D. Kakhki, and A. Jannesari. 2020. Predictive modeling for occupational safety outcomes and days away from work analysis in mining operations. *International Journal of Environmental Research and Public Health*, 17(19).

Zotova, E., A. García-Pablos, and M. Cuadros. 2021. Vicomtech at MEDDOPROF: Automatic Information Extraction and Disambiguation in Clinical Text. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings.*

A Annex 1: Supplementary Materials

More information on the MEDDOPROF Gold Standard's occupation typology and the results of all submitted runs can be found on the MEDDOPROF Supplementary Materials, available on Zenodo[10].

[10] https://doi.org/10.5281/zenodo.5086075