Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 381-392 recibido 09-07-2024 revisado 18-07-2024 aceptado 25-07-2024

Overview of GenoVarDis at IberLEF 2024: NER of Genomic Variants and Related Diseases in Spanish

Resumen de GenoVarDis en IberLEF 2024: NER de Variantes Genómicas y Enfermedades Relacionadas en Español

Marvin M. Agüero-Torales,[1,2] **Carlos Rodríguez Abellán,**[1]
Marta Carcajona Mata,[1] **Juan I. Díaz Hernández,**[1] **Mario Solís López,**[1]
Antonio Miranda-Escalada,[3] **Sergio López-Alvárez,**[1] **Jorge Mira Prats,**[1]
Carlos A. Castaño Moraga,[1] **David Vilares,**[4] **Luis Chiruzzo**[5]
[1]CoE of Data Intelligence, Fujitsu, Madrid, Spain
[2]Universidad de Granada, Granada, Spain
[3] LHF Labs, Bizkaia, Spain
[4] Universidade da Coruña, CITIC, A Coruña, Spain
[5]Universidad de la República, Montevideo, Uruguay
maguero@correo.ugr.es, {carlos.rodriguezabellan, marta.carcajonamata,
juanignacio.diazhernandez, mario.solislopez}@fujitsu.com, antoniomiresc@gmail.com,
{sergio.lopezalvarez, jorge.miraprats, carlosalberto.castanomoraga}@fujitsu.com,
david.vilares@udc.es, luischir@fing.edu.uy

Abstract: We present the first shared task for Named Entity Recognition (NER) in Spanish written scientific literature about genomic variants, genes, and its associated diseases and symptoms, GenoVarDis at IberLEF 2024 campaign. The challenge consisted on identifying entities related to genomic variants. We annotated a corpus of 633 abstracts extracted from PubMed articles, with the information for the tasks. Seven teams took part in the evaluation phase, obtaining in general good results for the task.

Keywords: NER, Genomic variants, Genetic diseases, Spanish.

Resumen: Presentamos la primera competencia de Reconocimiento de Entidades Nombradas (NER) en literatura científica escrita en español sobre variantes genómicas, genes y sus enfermedades y síntomas asociados, GenoVarDis en la campaña IberLEF 2024. El reto consistía en identificar entidades relacionadas con variantes. Se anotó un corpus de 633 resúmenes extraídos de artículos de PubMed, con la información para las tareas. Siete equipos participaron en la fase de evaluación, obteniendo en general buenos resultados para la tarea.

Palabras clave: NER, Variantes genómicas, Enfermedades genéticas, Español.

1 Introduction

This paper presents the GenoVarDis shared task for Named Entity Recognition (NER) in Spanish written scientific literature about genomic variants, genes, and its associated diseases and symptoms at IberLEF 2024 (Chiruzzo, Jiménez-Zafra, and Rangel, 2024).

Genomic variants are alterations in the DNA sequence that can influence the function and regulation of genes, potentially leading to various diseases and conditions (1000 Genomes Project Consortium et al., 2010). Accurate identification and classification of these variants are crucial for understanding their roles in health and disease, facilitating advances in personalized medicine and genomic research (Wei et al., 2022).

Variants recognition is important because it enables the detection of mutations that may cause or contribute to the development of diseases. This process aids in the diagnosis, treatment, and prevention of genetic disorders, improving patient outcomes and contributing to the field of genomics (Chen et al., 2023). On the other hand, related enti-

ISSN 1135-5948 DOI 10.26342/2024-73-29

ties such as diseases, symptoms and genes are relevant on the basis that they provide context and help in understanding the broader implications of genomic variants. Recognizing these entities allows for the construction of detailed genetic and biomedical databases, which are essential for research and clinical decision-making (Walsh et al., 2024). It also facilitates the linking of genetic data with phenotypic information, contributing to the development of comprehensive genetic models and improving our understanding of disease mechanisms (Nuzzo, Riva, and Bellazzi, 2009).

In this work, we aim to enhance the diversity of tasks at IberLEF 2024 (Chiruzzo, Jiménez-Zafra, and Rangel, 2024) and foster advancements in Spanish biomedical text processing. Effective NER systems can streamline the extraction of relevant information from vast amounts of biomedical literature, making critical data more accessible to researchers and clinicians (Cho and Lee, 2019). This task contributes to addressing the scarcity of resources in this specific domain, providing participants with an opportunity to advance research in NER. These PubMed[1] abstracts were annotated by the organisers to create the shared task corpus, which is available online.[2]

The paper is structured as follows: in Section 2 we mention the previous work on NER in genomic variants; in Section 3 we detail the task that make up this shared task; in Section 4 we present the corpus created for this work; in Section 5 we present the competition, the systems built by the participant teams and their results; and in Section 6 we detail the conclusions of this work.

2 Background

There have been some competitions focusing on NER in the biomedical domain, such as the BioCreative challenges (Rezarta et al., 2023) and the BioNLP shared task (Demner-fushman, Ananiadou, and Cohen, 2023). These competitions have traditionally focused on recognizing entities such as genes, proteins, diseases, and chemicals, mainly in English texts. As far as we know, this is the first time genomic variants is present as subject of a NER competition, and also the first dataset with this kind of annotations built exclusively for the Spanish language.

Variant NER datasets and systems are almost nonexistent, even in English, for example, `tmVar3`[3] (Wei et al., 2022) with just 500 documents or `BERN2`[4] (Sung et al., 2022), which uses tmVar2 (Wei et al., 2018) for variant detection, with only 158 documents. Another small English corpus is `Variome`[5] with ten PubMed articles, which had a broader definition of genetic variant than some prior works and also had relations (Verspoor et al., 2013; Verspoor et al., 2016). Cheng, Tan, and Wei (2020) work specifically addresses genomic variant recognition in biomedical literature using an end-to-end deep learning approach. It highlights the challenges of low-resource domains and linguistic heterogeneity in genomic variant entities. While most tools employ regular expressions based on the Human Genome Variation Society (HGVS) nomenclature,[6] the existing landscape faces limitations in recognizing diverse variant types (Yepes and Verspoor, 2014; Lee, Wei, and Lu, 2020). The understanding of genetic diseases relies heavily on the automated gathering and synthesis of published knowledge about sequence variants from scientific literature (Wei et al., 2022).

These previous studies have explored the identification of genomic variants in biomedical texts, highlighting the importance of this task in understanding genetic information. These works primarily focused on English-language resources, developing models and tools to recognize genetic variants and associated entities in scientific literature and clinical records.

NER on genomic variants can be considered a low resource task because, despite having an important presence in the scientific literature, electronic health records (EHR), and clinical histories, it does not have many open resources to work with, it is hard to annotate, and it has been mostly under-researched from the NLP perspective, even in English, which has the most resources available. Spanish, on

[1] https://pubmed.ncbi.nlm.nih.gov/

[2] https://codalab.lisn.upsaclay.fr/competitions/17733#participate-get_starting_kit

[3] https://github.com/ncbi/tmVar3

[4] https://github.com/dmis-lab/BERN2

[5] https://bitbucket.org/readbiomed/variome-corpus-data

[6] https://www.hgvs.org/

the other hand, even having some resources built for other biomedical and clinical domains, still has a long way to go for genomic variants and NLP.

3 Task

This task seek to expand the variety of tasks at IberLEF 2024 (Chiruzzo, Jiménez-Zafra, and Rangel, 2024) and encourage advancements in Spanish biomedical text processing. The GenoVarDis shared task contributes to addressing the scarcity of resources in this specific domain, providing participants with an opportunity to advance research in NER.

NER Given a text (sequence of tokens), identify the named entities as spans in the text and classify them according to one of those present in Table 1. Participants must find the beginning and the end of relevant mentions and classify them in the corresponding category.

Metric Participants were evaluated based on the accuracy of named entity recognition. The criterion for finding a named entity is exact match. Metrics included precision, recall, and F1 scores for the task (micro-averaged F-score is the primary metric). A prediction is successful if its span matches completely the Gold Standard annotation and has the same category.

Baselines We have compared every system to a baseline prediction (`Baseline-1`),[7] following Miranda-Escalada et al. (2021) baseline for the ProfNER shared task. The baseline for this task is a `Levenshtein` lexical lookup approach with a sliding window of varying length. This system serves the purpose of extracting information from a set of annotated documents and subsequently verifying whether these extracted annotations are present in a new set of documents (Agüero-Torales and Miranda-Escalada, 2024).

In addition, we proposed another system for the *Post-Evaluation phase* (`Baseline-2`, see § 5.1), which using OpenAI's `GPT-3.5 Turbo` model (OpenAI, 2024) through prompt engineering techniques customized for NER in the GenoVarDis corpus. This system employs iterative prompt refinement and optimizing to instruct the model in identifying and classifying the entities.

4 Corpus

The corpus comprises (i) the translation and manual curation of the documents of the `tmVar3` (Wei et al., 2022) annotations (composed by PubMed abstracts) and then we added the associated diseases and symptoms to the corpus; and (ii) the manual annotation of Spanish PubMed abstracts.

For (i), we use `GPT-3.5 Turbo` API (OpenAI, 2024) for the translation pipeline, which comprehends not only the abstracts themselves but also the entity names. In addition to the translation, for the entity projection, we use some regular expression (regex) to locate the translated entities within the translated abstract as the base corpus, prior to human curation. Meanwhile, for (ii) we downloaded the abstracts in May 2024 using the `PubMed Bio.Entrez` package.[8] The pipeline[9] used to download the abstracts is based on the library implemented in Miranda-Escalada et al. (2023). We kept only those articles in which the title and abstract were written in Spanish, after querying the API using a specific query. As in (i), we project PubTator3 API[10] entities detected and present in Table 1 to the annotators prior to the annotation process.

In the annotation process, we used the `brat` annotation tool[11] (Stenetorp et al., 2012) and three expert bioinformaticians took part. After the process was over we estimated the inter-annotator agreement (IAA) in the following way: we took randomly a subsample of 75 abstracts, 25 for each annotator that participated in the task. The three annotators completed the annotation of the 75 texts, and we compared the annotations as the pairwise agreement (intersection over union) of one annotator against another. The average obtained in this way was 0.78 (a substantial agreement) considering exact match (abstract, categories and offsets).

The data is split in approximately 70%-

[7] `https://github.com/mmaguero/genovardis-baseline`

[8] `https://biopython.org/docs/1.75/api/Bio.Entrez.html`

[9] `https://github.com/tonifuc3m/pubmed-parser`

[10] `https://www.ncbi.nlm.nih.gov/research/pubtator3/api`

[11] `https://brat.nlplab.org/introduction.html`

Name	Type	Example	Example (es)
DNAMutation	Variant on DNA sequence	c.1922G>A	c.1922G>A
		C-to-G transition was identified at nucleotide 857	la transición de C a G se identificó en el nucleótido 857
SNP	RS number	rs763780	rs763780
	COSMIC mutation	COSV53035892	COSV53035892
DNAAllele	Allele on DNA sequence	-218G	-218G
NucleotideChange/ BaseChange	Wild type and mutant	G > C, G/C	G > C, G/C
OtherMutation	Variant with insufficient information	306 base pair insertion	inserción de 306 pares de bases
		insertion introduced eight additional amino acids	la inserción introdujo ocho aminoácidos adicionales
Gene	Gene	ABCA1	ABCA1
Disease	Disease/Symptom	Congenital hypothyroidism	Hipotiroidismo congénito
		fever	fiebre
Transcript	Transcript ID	NM_015420.7, ENST00000413302	NM_015420.7, ENST00000413302

Table 1: Entity categories present in the corpus.

10%-20% for training, development (dev) and test sets. The Table 2 shows the characteristics of this corpus and it split, while Table 3 shows the entities distribution in each split.

As seen in the Table 2, the corpus contains more tmVar3's abstracts than Spanish PubMed set. The test data are slightly different than the dev and train data, because the last comprises the text translation and entities projection (from English to Spanish) and manual curation of the tmVar3's PubMed abstracts annotations (Wei et al., 2022) with their associated diseases and symptoms; while the test set, comprises the manual annotation of PubMed abstracts, originally written in Spanish (published between 2014 to 2024).

It is evident in Table 3, that the corpus is highly unbalanced, with a notable predominance of the `Disease` and `Gene` classes. Additionally, it is worth noting the scarcity of samples in the `Transcript` class, which appears only once across the training, dev, and test sets.

5 *Competition*

The competition ran between April 5 and June 9 (2024) on the CodaLab platform[12] (Pavao et al., 2022). A total of 35 users registered to participate. The number of participants that submitted results was approximately the 20% of the registered: five of them participated both in the development and the evaluation phase, and the other two participated only in the evaluation phase. Participant teams came both from the industry (2) and academia (5), and from different countries such as Spain, Mexico and Australia.

5.1 Phases

The competition consisted of three phases:

Development phase From April 5 to May 26. This phase started with the publication of the training and development sets. During this phase the participants could submit their predictions for the development set and get the correspondent score for the task. Each participant could make up to 200 submis-

[12] https://codalab.lisn.upsaclay.fr/competitions/17733

Texts	**tmVar3**		**PubMed-es**		**Total**	
	Count	**(%)**	**Count**	**(%)**	**Count**	**(%)**
Train	427	67.46	0	0.00	427	67.46
Dev	70	11.06	0	0.00	70	11.06
Test	0	0.00	136	21.48	136	21.48
Total	497	78.52	136	21.48	633	100

Table 2: Composition of the corpus.

Category	**Train**		**Dev**		**Test**		**Total**	
	Count	**(%)**	**Count**	**(%)**	**Count**	**(%)**	**Count**	**(%)**
Disease	4,028	49.13	588	44.11	1,433	68.21	6,049	52.00
Gene	3,093	37.72	550	41.26	514	24.46	4,157	35.73
DNAMutation	496	6.05	103	7.73	73	3.47	672	5.78
OtherMutation	271	3.31	53	3.98	22	1.05	346	2.97
SNP	120	1.46	15	1.13	42	2.00	177	1.52
DNAAllele	139	1.70	12	0.90	15	0.71	166	1.43
NucleotideChange/ BaseChange	51	0.62	11	0.83	1	0.05	63	0.54
Transcript	1	0.01	1	0.08	1	0.05	3	0.03
Total	8,199	70.48	1333	11.46	2,101	18.06	11,633	100

Table 3: Total number of category/entities in the split.

sions. There were 14 successful submissions.

Evaluation phase From May 27 to June 9. This phase started with the publication of the test set. In this phase the participants could submit the predictions of their final systems and get the correspondent score for the task. Each participant could make up to 10 submissions. There were 47 successful submissions.

Post-Evaluation phase From June 10 onward. This phase started after the end of the competition. The CodaLab page remains available for everyone who wants to test additional systems, download the training, development and test sets, and check the shared task information.

5.2 Systems description

We had seven participants in the evaluation phase, although we expected many participants would try prompting generative LLM (Large Language Model), these submissions are not among the best. Overall the followed approaches were notoriously diverse. We briefly describe those systems:

The Fujitsu Research of Europe team (`FRE`), Spain, **ander.martinez** (Martínez, 2024), highlights the importance of combining multiple techniques for robust NER performance, particularly in biomedical contexts with varied data distributions. The winning approach and solution in the GenoVarDis task involved fine-tuning pretrained Language Models (LMs), `bsc-bio-ehr-es` RoBERTA model (Carrino et al., 2022; Liu et al., 2019), using Conditional Random Fields (CRF) (Lafferty, McCallum, and Pereira, 2001), Byte-Pair Encoding dropout (BPE dropout) (Provilkov, Emelianenko, and Voita, 2020), and model ensemble. Firstly, it employs a token classification approach using the `IOB` schema, which structures labels for entities in text. Secondly, the team utilizes CRFs to model the transition probabilities between these labels, thereby enhancing the accuracy of predicting sequences of entities. Additionally, the team explores subword representation techniques through BPE, which generates subword units to effectively handle out-of-vocabulary words. To further boost model robustness, BPE dropout is applied to introduce randomness during the merging of subwords. Lastly, the study employs model ensembling by training five models with vary-

User	Team/Affiliation Country, A/I	F1	Precision	Recall
Development phase				
ander.martinez	FRE, Spain, I	**0.7403**	*0.7006*	**0.7847**
VictorMov	UGR, Spain, A	*0.6683*	**0.7094**	*0.6317*
orlandxrf	IIMASnlp, Mexico, A	0.4733	0.6286	0.3796
Baseline-1	-	0.4197	0.6504	0.3098
Antares-Amazel	BUAP, Mexico, A	0.3324	0.6478	0.2236
Milimeter98	RMIT-READ-BioMed, Australia, A	0.2356	0.2428	0.2288
Test phase				
ander.martinez	FRE, Spain, I	**0.8210**	**0.8223**	**0.8196**
VictorMov	UGR, Spain, A	*0.7935*	*0.7906*	*0.7963*
ELiRF-VRAIN	ELiRF-VRAIN, Spain, A	0.7349	0.7775	0.6968
Milimeter98	RMIT-READ-BioMed, Australia, A	0.5483	0.6108	0.4974
orlandxrf	IIMASnlp, Mexico, A	0.5301	0.7318	0.4155
GuillemGSubies	I	0.4283	0.4355	0.4212
Baseline-1	-	0.3194	0.5938	0.2185
Antares-Amazel	BUAP, Mexico, A	0.3009	0.6040	0.2004
Post-Evaluation phase				
Baseline-2	-	0.5129	0.4572	0.5840

Table 4: Results of the evaluation phase. Best result bolded, second best in italic. A/I stands for Academy/Industry.

ing initializations and employing a majority voting strategy to combine predictions, thus improving overall prediction stability.

Oliveros (2024a), **VictorMov**, presented a diverse exploration of NER systems[13] adapted for genomic variants and related diseases within biomedical texts in Spanish (Oliveros, 2024b). The team from the University of Granada (*UGR*), Spain, introduced three distinct models: `GPT-3.5 Turbo` (OpenAI, 2024), `roberta-base-biomedical-es` (Carrino et al., 2021), and `gliner_medium-v2.1` (Zaratiana et al., 2024). `GPT-3.5 Turbo`, known for its broad applicability and efficiency in natural language understanding tasks, was adapted for entity recognition through iterative prompt engineering. The RoBERTa, pretrained on a comprehensive Spanish biomedical corpus and fine-tuned for the NER task, demonstrated gradual improvement in precision, recall, and F1-score metrics across training epochs. GLiNER (medium size), a specialized NER model, was optimized for biomedical text, excelled with its bidirectional transformer architecture, consistently outperforming RoBERTa and `GPT-3.5 Turbo` in identifying and classifying entities. Despite challenges such as class imbalance in entity types and varying model architectures impacting training setups, the team navigated these complexities to achieve competitive performance in the task, underscoring the models' adaptability and effectiveness in the biomedical domain.

In their participation, Marco, Segarra, and Hurtado (2024), **ELiRF-VRAIN**, from Language Engineering and Pattern Recognition (*ELiRF*) group of the Valencian Research Institute for Artificial Intelligence (*VRAIN*) at *Universitat Politècnica de València*, Spain, utilizing two base pre-trained models, specifically `bsc-bio-ehr-es` RoBERTa (Carrino et al., 2022) and `CLIN-X-ES` XLM-RoBERTa (Lange et al., 2021; Conneau et al., 2020), they adopt the `IOB2` labeling scheme for entity classification and boundary detection. Fine-tuning these models on the GenoVarDis corpus involved optimizing hyperparameters such as epochs, learning rates, and batch sizes using `Optuna` (Akiba et al., 2019) for micro-F1 score maximization on validation data. Their approach includes four systems, integrating different pre-trained models and hyperparameter configurations, culminating in competitive re-

[13] `https://github.com/Victor-mov/GenoVarDis`

sults reflective of their systematic evaluation and model selection strategies.

The system[14] of the RMIT University team (`RMIT-READ-BioMed`, Australia, **Milimeter98**), Kodikara and Verspoor (2024a), focused on NER using a generative LLM, specifically OpenAI's `GPT-3.5 Turbo` (OpenAI, 2024). Their approach explores cross-linguistic settings by providing English-language instructions alongside Spanish-language texts, comparing this with within-language settings. They experiment with various prompting strategies, including zero-shot and few-shot learning paradigms, finding that the best performance, is achieved under the few-shot learning paradigm using English-language instructions. Despite not achieving top results in the competition, their exhaustive work provides insights into the challenges and limitations of using LLMs for NER in languages other than English, highlighting the effectiveness of annotated guidelines and prompting strategies in enhancing model performance (Kodikara and Verspoor, 2024b).

Ramos-Flores, Gómez-Adorno, and Galán-Vásquez (2024), **orlandxrf**, took part as the `IIMASnlp` team from IIMAS-UNAM (Applied Mathematics and Systems Research Institute - National Autonomous University of Mexico), Mexico. They developed a system for the GenoVarDis task involving NER using a combination of CRF, Bi-directional Long Short-Term Memory (Bi-LSTM) networks (Hochreiter and Schmidhuber, 1997; Huang, Xu, and Yu, 2015), fine-tuned `roberta-base-biomedical-clinical-es` (Carrino et al., 2021), and a zero-shot approach. They enhanced their training dataset with a data augmentation technique using a quantized 4-bit version of `LLaMA3-8b` (AI@Meta, 2024), significantly increasing the data volume. Their methodology included pre-processing, training several supervised models, and experimenting with zero-shot NER using 4-bit `LLaMA3-8b`. They utilized a post-processing step that combined predictions from their best-trained CRF model with the zero-shot results, applying lexicons and regular expressions to refine entity extraction. This multifaceted approach, combining CRF and zero-shot methods with a robust post-processing phase, achieved their highest performance scores in the task (5th position).

The systems submitted by **GuillemG-Subies** for our task utilized various LLMs including `large` XLM-RoBERTa (Conneau et al., 2020), `RigoBERTa-2.0` DeBERTa (Vaca Serrano et al., 2022; He et al., 2021), `bsc-bio-ehr-es` RoBERTa (Carrino et al., 2022), and `roberta-large-bne` (Fandiño et al., 2022) to perform NER. The approach includes multiple submissions to the competition platform Codalab, experimenting with different model configurations with and without CRF integration and prefer or no the first entity-label predicted. The submissions reveal a systematic exploration of model performance, with models being iteratively refined and evaluated to identify the optimal configurations for NER tasks. The best model for this participant was a RigoBERTa without CRF.

Lezama-Sánchez and Tovar Vidal (2024), **Antares-Amazel**, from *Benemerita Universidad Autonoma de Puebla* (*BUAP*), Mexico, focused on NER using LSTM neural networks. Starting with rigorous text preprocessing steps to normalize and clean data, including tasks like accent removal and punctuation elimination, the authors constructed a sequential neural network model. The architecture features two Bi-LSTM layers with different configurations to capture contextual information bidirectionally within textual sequences, complemented by dense layers with ReLU activation for nonlinear learning and a softmax output layer for multiclass entity classification. Te results obtained with this methodology were a not the best, but the team indicates it was a first attempt, and probably improving the architecture and tuning the hyperparameters would give better results. Furthermore, in this type of architecture, removing some tokens like stopwords and punctuation could have negative impact in performance, because the network would rely on these functional tokens to understand the overall syntax of the sentence (which is important for NER).

5.3 Results

Table 4 shows results of the development phase and the final results of the evaluation phase. We also had an additional submission

[14] `https://github.com/Milindi-Kodikara/RMIT-READ-BioMed`

by our baseline-2 for the post-evaluation phase that ranks next to the top 5 of the systems of the evaluation phase.

In general, most of the participants got very good results for the task, beating the baselines by a good margin. Only one team could not beat the Baseline-1 and two the Baseline-2.

The categories where the participants got the best performance for the task were Disease, Gene, and DNAMutation, being the second one an easy class to predict, containing mostly nomenclatures. Conversely, the hardest classes to classify were Transcript, NucleotideChange-BaseChange and DNAAllele, which also had the fewest examples (see Table 3).

The terms in the OtherMutation category are notably longer than the average length of other entities. As indicated in Table 1, these terms typically consist of detailed descriptions of mutations. In contrast, mentions in the Transcript and SNP categories are single words, adhering to a consistent pattern that allows for extraction using regular expressions.

The GenoVarDis shared task saw diverse approaches to NER in genomic variants and related diseases, showcasing a variety of strategies and models. The FRE team won the competition by fine-tuning pretrained LLMs like the bsc-bio-ehr-es RoBERTa model and integrating CRF, BPE dropout, and model ensemble techniques. The *UGR* team (**VictorMov**) explored three distinct models: GPT-3.5 Turbo, roberta-base-biomedical-es, and gliner_medium-v2.1; each demonstrating varying degrees of effectiveness, with GLiNER performing best. The ELiRF-VRAIN team utilized bsc-bio-ehr-es RoBERTa and CLIN-X-ES XLM-RoBERTa models, optimizing hyperparameters with Optuna, obtained the best results with the second ones. The RMIT-READ-BioMed team focused on using GPT-3.5 Turbo in cross-linguistic settings, finding the few-shot learning paradigm most effective. The IIMASnlp team combined CRF, Bi-LSTM networks, fine-tuned roberta-base-biomedical-clinical-es, and a zero-shot approach with data augmentation using LLaMA3-8b. **GuillemG-Subies** employed various models with RigoBERTa-2.0 without CRF performing best. The *BUAP* team (**Antares-Amazel**) focused on Bi-LSTM neural networks but noted their results could improve with further tuning and architecture adjustments.

The results achieved for the teams highlighted the effectiveness of combining multiple NER techniques to handle the complexities of biomedical texts. The winning approach by the FRE team demonstrated the benefits of integration and model ensembling. Other teams, like those from the RMIT-READ-BioMed and IIMASnlp, showcased the adaptability and robustness of their models through diverse methodologies, including prompt engineering and data augmentation. Despite varying levels of success, all participants contributed valuable insights into the challenges and strategies of NER in genomic variants and related diseases. Several teams identifying areas for future improvement. This collective effort advances the field of biomedical NER, providing a foundation for continued innovation and optimization.

6 Conclusions

In this work, we presented the GenoVarDis shared task for NER of genomic variants and related diseases in Spanish at IberLEF 2024, the first task of its kind. In addition, the GenoVarDis shared task can be used as template for future shared tasks on the recognition of genomic variants in other languages, beyond English.

GenoVarDis, which focuses on genomic variants and related diseases in Spanish, has revealed that participants tend to exhibit low precision but high recall in their systems. This suggests that while many relevant entities are being identified, there is a significant number of false positives, which can hurt the accuracy and reliability of the results. To enhance the precision of our corpus, future iterations could incorporate more strict annotation guidelines and increased diversity in the training data to cover underrepresented entities more effectively. Additionally, since the majority classes (Disease and Gene) constitute 90% of the corpus, there is a risk of participants optimizing their systems primarily for these abundant classes, potentially neglecting more critical yet less frequent entities such as genomic mutations. To counter this, we could implement a balanced sampling strategy or use weighted metrics that emphasize the importance of *rare* but significant classes. This approach would encourage participants to develop systems that are more robust and capable of accurately identifying a broader range of relevant entities, thereby improving the overall utility and reliability of the dataset.

Seven participants submitted their predictions for the evaluation phase, with the FRE team submission being the one with the best performance across all the metrics, winning the competition. GenoVarDis has aroused interest from both academia and industry. Interestingly, a team from a non-Spanish speaking country participated in this task. A couple of participant systems reached high performances. However, the detection and classification of genomic variants, genes and related diseases and symptoms data can still be improved. We hope this work contributes to spark interest in the detailed analysis of genomic variants in non-English contexts, encouraging further research and development in this crucial area.

Acknowledgments

Our thanks to all GenoVarDis shared task participants, as well as many thanks to the organizers of IberLEF 2024 for creating the perfect ecosystem for shared tasks.

References

1000 Genomes Project Consortium et al. 2010. A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061.

Agüero-Torales, M. M. and A. Miranda-Escalada. 2024. GenoVarDis-ST Baseline 1 - Lookup. https://github.com/mmaguero/genovardis-baseline.

AI@Meta. 2024. Llama 3 model card. *GitHub repository.*

Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Carrino, C. P., J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, and M. Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario.

Carrino, C. P., J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, and M. Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, editors, *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland, May. Association for Computational Linguistics.

Chen, E., F. M. Facio, K. W. Aradhya, S. Rojahn, K. E. Hatchell, S. Aguilar, K. Ouyang, S. Saitta, A. K. Hanson-Kwan, N. N. Capurro, E. Takamine, S. S. Jamuar, D. McKnight, B. Johnson, and S. Aradhya. 2023. Rates and Classification of Variants of Uncertain Significance in Hereditary Disease Genetic Testing. *JAMA Network Open*, 6(10):e2339571–e2339571, 10.

Cheng, C., F. Tan, and Z. Wei. 2020. Deepvar: an end-to-end deep learning approach for genomic variant recognition in biomedical literature. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 598–605.

Chiruzzo, L., S. M. Jiménez-Zafra, and F. Rangel. 2024. Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org.*

Cho, H. and H. Lee. 2019. Biomedical named entity recognition using deep neural networks with contextual information. *BMC bioinformatics*, 20:1–11.

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 D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Demner-fushman, D., S. Ananiadou, and K. Cohen, editors. 2023. *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Toronto, Canada, July. Association for Computational Linguistics.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

He, P., X. Liu, J. Gao, and W. Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Hochreiter, S. and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11.

Huang, Z., W. Xu, and K. Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Kodikara, M. and K. Verspoor. 2024a. Effectiveness of Cross-linguistic Extraction of Genetic Information using Generative Large Language Models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEURWS.org*, Valladolid, Spain, September.

Kodikara, M. and K. Verspoor. 2024b. The RMIT University system for NER of genetic entities in biomedical literature for the GenoVarDis shared task at IberLEF 2024. `https://github.com/Milindi-Kodikara/RMIT-READ-BioMed`.

Lafferty, J. D., A. McCallum, and F. C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lange, L., H. Adel, J. Strotgen, and D. Klakow. 2021. Clin-x: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain. *Bioinformatics*.

Lee, K., C.-H. Wei, and Z. Lu. 2020. Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature. *Briefings in Bioinformatics*, 22(3):bbaa142, 08.

Lezama-Sánchez, A. L. and M. Tovar Vidal. 2024. Named Entity Recognition in Scientific Texts Using Long Short-Term Memory Networks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEURWS.org*, Valladolid, Spain, September.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Marco, P., E. Segarra, and L.-F. Hurtado. 2024. ELiRF at GenoVarDis Task: NER in Genomic Variants and related Diseases. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEURWS.org*, Valladolid, Spain, September.

Martínez, A. 2024. FRE at GenoVarDis: A sane approach to Disease and Genomic Variant NER. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEURWS.org*, Valladolid, Spain, September.

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 A. Magge, A. Klein, A. Miranda-Escalada, M. A. Al-garadi, I. Alimova, Z. Miftahutdinov, E. Farre-Maduell, S. L. Lopez, I. Flores, K. O'Connor, D. Weissenbacher, E. Tutubalina, A. Sarker, J. M. Banda, M. Krallinger, and G. Gonzalez-Hernandez, editors, *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 13–20, Mexico City, Mexico, June. Association for Computational Linguistics.

Miranda-Escalada, A., F. Mehryary, J. Luoma, D. Estrada-Zavala, L. Gasco, S. Pyysalo, A. Valencia, and M. Krallinger. 2023. Overview of DrugProt task at BioCreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical–protein relations. *Database*, 2023:baad080, 11.

Nuzzo, A., A. Riva, and R. Bellazzi. 2009. Phenotypic and genotypic data integration and exploration through a web-service architecture. *BMC bioinformatics*, 10:1–11.

Oliveros, V. M. 2024a. GenoVarDis@IberLEF2024: Automatic Genomic Variants and Related Diseases using Named Entity Recognition with Large Language Models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEURWS.org*, Valladolid, Spain, September.

Oliveros, V. M. 2024b. GenoVarDis@IberLEF2024: Automatic Genomic Variants and Related Diseases using Named Entity Recognition with Large Language Models. `https://github.com/Victor-mov/GenoVarDis`.

OpenAI. 2024. OpenAI GPT-3.5 Turbo API [gpt-3.5-turbo:0301]. Accessed: 2024.

Pavao, A., I. Guyon, A.-C. Letournel, X. Baró, H. Escalante, S. Escalera, T. Thomas, and Z. Xu. 2022. Codalab competitions: An open source platform to organize scientific challenges. *Technical report.*

Provilkov, I., D. Emelianenko, and E. Voita. 2020. BPE-dropout: Simple and effective subword regularization. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online, July. Association for Computational Linguistics.

Ramos-Flores, O., H. Gómez-Adorno, and E. Galán-Vásquez. 2024. IMASnlp at GenoVarDis Task: Exploring Zero-shot and CRF Approaches to NER Task in Genomic Variants and Related Diseases. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEURWS.org*, Valladolid, Spain, September.

Rezarta, I., C. Arighi, I. Campbell, G. Gonzalez-Hernandez, L. Hirschman, M. Krallinger, S. Lima-López, D. Weissenbacher, and Z. Lu, editors. 2023. *Proceedings of the Biocreative VIII Challenge and Workshop: Curation and Evaluation in the Era of Generative Models*, New Orleans, USA. Zenodo.

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 F. Segond, editor, *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April. Association for Computational Linguistics.

Sung, M., M. Jeong, Y. Choi, D. Kim, J. Lee, and J. Kang. 2022. BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, 38(20):4837–4839, 09.

Vaca Serrano, A., G. García Subies, H. Montoro Zamorano, N. Aldama García, D. Samy, D. Betancur Sánchez, A. Moreno-Sandoval, M. Guerrero Nieto, and Á. Barbero Jiménez. 2022. RigoBERTa: A State-of-the-Art Language Model For Spanish. *ArXiv*, abs/2205.10233.

Verspoor, K., A. Jimeno Yepes, L. Cavedon, T. McIntosh, A. Herten-Crabb,

Z. Thomas, and J.-P. Plazzer. 2013. Annotating the biomedical literature for the human variome. *Database*, 2013:bat019, 04.

Verspoor, K. M., G. E. Heo, K. Y. Kang, and M. Song. 2016. Establishing a baseline for literature mining human genetic variants and their relationships to disease cohorts. *BMC medical informatics and decision making*, 16:37–47.

Walsh, N., A. Cooper, A. Dockery, and J. J. O'Byrne. 2024. Variant reclassification and clinical implications. *Journal of Medical Genetics*, 61(3):207–211.

Wei, C.-H., A. Allot, K. Riehle, A. Milosavljevic, and Z. Lu. 2022. tmVar 3.0: an improved variant concept recognition and normalization tool. *Bioinformatics*, 38(18):4449–4451, 07.

Wei, C.-H., L. Phan, J. Feltz, R. Maiti, T. Hefferon, and Z. Lu. 2018. tmvar 2.0: integrating genomic variant information from literature with dbsnp and clinvar for precision medicine. *Bioinformatics*, 34(1):80–87.

Yepes, A. J. and K. Verspoor. 2014. Mutation extraction tools can be combined for robust recognition of genetic variants in the literature. *F1000Research*, 3.

Zaratiana, U., N. Tomeh, P. Holat, and T. Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico, June. Association for Computational Linguistics.