ELiRF at GenoVarDis Task: NER in Genomic Variants and related Diseases

Pere Marco[1,*,†], Encarna Segarra[1,2,†] and Lluís-Felip Hurtado[1,†]

[1]VRAIN: Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain

[2]ValgrAI: Valencian Graduate School and Research Network of Artificial Intelligence, Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain

Abstract

In this work, we present our participation in GenoVarDis competition, the first edition of a competition of Named Entity Recognition in Spanish scientific literature about genomic variants, genes, and their associated diseases. Our approach consists of the use of pre-trained models specialized in the clinical domain fine-tuned for the Named Entity Recognition task. We present four systems that result from different base pre-trained models and different strategies for best model selection. Our systems have achieved competitive results compared to the models presented in the GenoVarDis competition.

Keywords

Natural Language Processing, Spanish Clinical NER, Token Classification, Transformers-based Models

1. Introduction

The Named Entity Recognition (NER) of genomic variants and Diseases task is very useful in the clinical setting. It can be used to locate texts in which certain diseases or genetic variants are discussed. In addition, we can use the entities detected to normalize specialized topics and expressions of the document and establish relationships between the contents of biomedical texts. It can also help with other tasks such as information retrieval and text classification of medical texts.

Nowadays, we can find some relevant systems such as tmVar3 [1] and BERN2[2]. Those models are used to find biomedical entities. Regarding tmVar3, it is possible to obtain genetic variants in abstracts and medical articles obtained from PubMed. We can find an example of the applicability of this model in the web app Pubtator, where tmVar3 is used along with other specialized models to perform NER tasks. BERN2 is used to identify drugs and diseases in texts, it allows multiple Named Entity Recognition and Normalization tasks on large biomedical texts in a fast and efficient way.

Recently, different models have been presented in Spanish for NER in the clinical setting. One of them is the RoBERTa-based model recently presented by Carrino et al. [3], which is trained from scratch using two different corpora with biomedical and Electronic Health Records (EHR) documents. Authors fine-tuned models for different NER tasks and compared them with other generic models and domain-specific models. The results showed the competitiveness of their model for clinical Natural Language Processing (NLP) applications. During the same period, the CLIN-X [4] models were published. In this work, models were presented, both in Spanish and English, for the extraction of information from clinical texts. These models were created from XLM-RoBERTA and show robust results in both general domain and specific domain NER tasks.

In our work, we approach the task as a classical NER task using the IOB2 scheme, where each sample has associated a sequence of labels, one assigned to each word. To implement our system, we take as a starting point two pre-trained models. For fine-tuning, we made a hyperparameter search using

IberLEF 2024, September 2024, Valladolid, Spain

*Corresponding author.

†These authors contributed equally.

pmarco@vrain.upv.es (P. Marco); esegarra@dsic.upv.es (E. Segarra); lhurtado@dsic.upv.es (L. Hurtado)

0009-0003-7026-3543 (P. Marco); 0000-0002-5890-8957 (E. Segarra); 0000-0002-1877-0455 (L. Hurtado)

the Optuna optimization framework and we used micro-F1 for selecting the best epoch. We present four systems that result from different base pre-trained models and different strategies for best model selection.

The rest of the paper is organized as follows. In Section 2, we present the GenoVarDis task. In Sections 3 and 4, we explain the dataset and the evaluation metrics, we also describe the proposed systems. Section 5 presents the results and their analysis. Finally, Section 6 indicates some conclusions and possible ways to continue this work.

2. Task Description

Named Entity Recognition is one of the tasks that has generated most interest in the field of NLP. Entity detection has proven to be useful in a wide range of research areas, both as a process in itself and as an intermediate step in the development of other activities such as text classification or Question Answering. In the GenoVArDis [5] task, presented in the framework of the IberLEF 2024 [6] workshop, the goal is to be able to locate DNA mutations and variant-related entities (such as genes, diseases and symptoms) in medical texts in Spanish. The texts consist of abstracts from different sources, mainly, translated from English and curated by human experts. We can see an example of these texts in Figure 1, where the entities are highlighted with different colors, depending on the class they belong to.

The objective of this task is to provide the exact beginning and end of the entities that we can find in the provided texts, that is, to identify the named entities as spans in the text. These spans have no overlaps or discontinuities. Each detected span is classified into one of the following 8 entity classes: *DNAMutation, SNP, DNAAllele, NucleotideChange/BaseChange, OtherMutation, Gene, Disease*, and *Transcript*.

Figure 1: Labeled text from GenoVarDis corpus.

12626442|t| Deficiencia total de C4B debido a la eliminación y conversión génica en un paciente con infecciones graves.
12626442|a|Las deficiencias de los componentes tempranos de la vía clásica del complemento afectan las acciones de la inmunidad innata y humoral y pueden aumentar la susceptibilidad a las infecciones. Hemos estudiado la base genética de la deficiencia total de C4B en un paciente finlandés con meningitis recurrente, fístulas crónicas y abscesos. El cromosoma materno llevaba una eliminación de cuatro genes, incluyendo el gen C4B, y se encontró una conversión del gen C4B al gen C4A en el cromosoma paterno, lo que resultó en una deficiencia completa de C4B. En el gen C4A convertido, el análisis de mutaciones no reveló cambios de aminoácidos o mutaciones prominentes, pero se encontró un gran número de variaciones nucleotídicas. Además, el paciente era heterocigoto para la deficiencia estructural de la lectina de unión a manano (MBL), lo que se asoció con niveles medios de MBL en suero. Nuestros datos proporcionan nueva información sobre la inestabilidad genética de la región del gen C4 y sobre la asociación de la deficiencia homocigota de C4B y el genotipo variante de MBL con una mayor susceptibilidad a infecciones recurrentes y crónicas. Es importante destacar que la terapia plasmática indujo una cura clínica rápida con efectos a largo plazo.
Disease
Gene

3. The Dataset and Evaluation Metrics

The corpus provided by the GenoVarDis task organizers is composed of medical abstracts obtained from two processes. Many of the texts that make up the dataset come from the tmVar3.0 [1] corpus. This corpus is composed of annotated abstracts that have been extracted from PubMed in English. Part of the GenovarDis corpus consists of some of these texts that have been translated from English and curated by human experts. The remaining texts consist of abstracts in Spanish, obtained from PubMed or SciELO, that were manually annotated. As a result, the competition provides an annotated corpus of 633 samples in Spanish. Table 1 shows the distribution of the partitions and relevant statistics of the corpus. A more detailed description of the data can be found in [5].

Table 1
Dataset statistics.

	Documents		Entities	
Partition	Number	(% Total)	Number	(% Total)
Training	427	(67.46%)	8,199	(70.48%)
Validation	70	(11.06%)	1,333	(11.46%)
Test	136	(21.48%)	2,101	(18.06%)
Total	633	(100.00%)	11,633	(100.00%)

The dataset samples have been labeled to locate entities of 8 different classes. Each entity can only belong to a single class. Furthermore, the entities have no overlap, neither partial nor complete. There are also no discontinuous entities or dependent entities between sentences. Table 2 presents the distribution of entities by entity class of each of the corpus partitions ordered by frequency in the training set. It can be seen that it is a highly unbalanced corpus, with a clear predominance of the *Disease* and *Gene* classes. We can also highlight the scarcity of samples of the *Transcript* class, which has only a single occurrence in the training, validation, and test sets. The organization has set the micro-F1 scores for the evaluation of the systems participating in the GenoVarDis competition. To consider a sample as correct, both the start character and the end character of the label must match, i.e., an exact match of each entity is required to positively count the score.

Table 2
Distribution of Entities (Ent.) by corpus partition. Entity classes are ordered by frequency in the training set.

	Training		Validation		Test	
Entity Class	# Ent.	(%)	# Ent.	(%)	# Ent.	(%)
Disease	4,028	49.13%	588	44.11%	1,433	68.21%
Gene	3,093	37.72%	550	41.26%	517	24.46%
DNAMutation	496	6.05%	103	7.73%	73	3.47%
OtherMutation	271	3.31%	53	3.98%	22	1.05%
DNAAllele	139	1.70%	12	0.90%	15	0.71%
SNP	120	1.46%	15	1.13%	42	2.00%
NucleotideChange/ BaseChange	51	0.62%	11	0.83%	1	0.05%
Transcript	1	0.05%	1	0.08%	1	0.05%
Total	8,199	100.00%	1,333	100.00%	2,101	100.00%

4. System Description

4.1. Overview of the System

We approached the problem as a classical NER problem. In this way, each sample has associated a sequence of labels indicating the class of entity to which each word belongs. Since the objective is not only to classify the tokens but also to delimit where each entity starts and where it ends, we decided to use the IOB2 labeling scheme. In this way, in case of finding two consecutive entities of the same class, we will be able to determine if the entities have been correctly fragmented.

To implement our systems we have decided to use as a basis two pre-trained models: the RoBERTa type model of Carrino [3], *PlanTL-GOB-ES/bsc-bio-ehr-es*, and the XLM-RoBERTa model of Lange [4], *CLIN-X-ES*.

For pre-training Carrino's model, the authors used two different corpora in Spanish: an EHR corpus and a biomedical one. The EHR corpus contains 95M tokens from clinical documents, including x-ray reports, discharge reports, and clinical course notes. The biomedical corpus includes Spanish data from a variety of sources for a total of 2,5M documents. This model was pre-trained using a RoBERTa-based architecture from scratch.

In the case of CLIN-X, they developed XLM-RoBERTa models for the clinical domain using English and Spanish corpora. In our work, we used the Spanish model. To steer this model towards the Spanish clinical domain the authors used documents from the Scielo archive and the MeSpEn resources [7], with a total of 790MB of highly specific data.

Both base models were fine-tuned using the IOB2 scheme. Since after tokenizing all the samples were shorter than the models' input size, any text preprocessing was necessary before fine-tuning the models. In the case of the system trained from Carrino's model, we used two different metrics to evaluate the performance of the fine-tuning process. The first metric acted like an IO scheme, where the separation between consecutive entities of the same class was ignored. This is the strategy used in System 1. For System 2 we used Carrio's pre-trained model, but in this case, the limits of the consecutive entities were relevant for model evaluation. Using the same scheme than System 2 we develop System 3, but in this case, we used CLIN-X-ES as the base model. Finally, we presented System 4 as a result of combining the two best models obtained, System 2 and System 3. In this combination, the selection of labels in case of overlap has been resolved taking into account the model with greater Precision.

Table 3
Different combinations of hyperparameter search strategy for the Systems 1, 2, and 3.

Parameter	System 1	System 2	System 3
Epochs	39	50	41
Learning rate	4.16e-05	3.29e-05	4,54e-05
Batch size	4	2	4
Optimizer	AdamW	AdamW	AdamW
Gradient accum. steps	8	8	2
Weight decay	5.42e-03	8.18e-03	4.02e-03
Lr scheduler	Linear	Linear	Constant

4.2. Hyperparameter Optimization

In the fine-tuning process we used Optuna [8], for the hyperparameter search based on the micro-F1 results on the validation set. The parameters indicated to Optuna were: the number of epochs, from 20

to 60; the learning rate, from 1e-3 to 1e-7; batch size, among 2, 4and 8; gradient accumulation steps, among 2, 4, 8, 16, and 32; weight decay, from 1e-5 to 1e-2; and learning rate schedule type, between constant and linear. The remaining parameters were kept at the Hugging Face Trainer default values (see https://huggingface.co/docs/transformers/main_classes/trainer#transformers.TrainingArguments).

We present four systems obtained using the training parameters shown in Table 3: System 1 and 2, developed from RoBERTa-based pre-trained model by Carrino, System 3, trained using the CLIN-X-ES model, and finally System 4, which is a combination of the two best previous systems based on their test set performance. Training values are not presented for System 4 since it is the combination of two models trained separately. This last system prioritizes the labels obtained by the CLIN-X-based model since it has obtained the best Precision results.

5. Experimental Results and Discussion

Table 4 shows the results of the four systems on the test set in terms F1, Precision, and Recall using a micro averaging schema. The best model is System 3, generated from the CLIN-X-ES pre-trained model. It is remarkable the great difference between the Precision and Recall of this system. Although a high percentage of the labels it generates are correct, it presents a worse performance in generating the true positive labels. The second best system in terms of micro-F1 is the system trained from Carrino's model, with strict separation of entities. Finally, as expected, System 4, which combines the two best systems, presents a higher Recall value, since it gets more correct labels, but when committing the errors of both systems, it drops significantly in Precision, obtaining the worst micro-F1 result.

Table 4
Results of the four systems on the test set.

	F1	Precision	Recall
System 1	0.7093	0.7051	0.7135
System 2	0.7240	0.7334	0.7149
System 3	0.7349	0.7775	0.6968
System 4	0.6659	0.6204	0.7187

6. Conclusions and Future Work

In this paper, we present our approach for the automatic recognition of genomic variants and related diseases entities in Spanish medical texts. Our approach consists of the use of pre-trained models specialized in the clinical domain fine-tuned for the NER task. Our systems have achieved competitive results compared to the models presented in the GenoVarDis competition.

In our work, we present four systems obtained from the use of different pre-trained models and the combination of predictions from several models. For fine-tuning, we used the IOB2 scheme and we made a hyperparameter search using the Optuna optimization framework and micro-F1 metric.

The work done presents several possible ways of improvement. The use of CRFs in combination with the pre-trained models or the use of Data Augmentation for the NER task would be the most direct and interesting extensions for the improvement of the systems. Another way of work would be to investigate the combination of model predictions, either from our own trained models or combining tools such as BERN2.

7. Ethics Statement

The Carrino's and CLIN-X pre-trained models used as base models for the task are obtained from HuggingFace models hub, under the Apache License 2.0 and CC-BY 4.0 license respectively.

Acknowledgments

This work is partially supported by MCIN/AEI/10.13039/501100011033 and ERDF A way of making Europe under grant PID2021-126061OB-C41. It is also partially supported by the Generalitat Valenciana under the GVA-Predoctoral Research Grant (CIACIF/2022/234) and CIPROM/2021/023 project.

References

[1] C.-H. Wei, A. Allot, K. Riehle, A. Milosavljevic, Z. Lu, tmvar 3.0: an improved variant concept recognition and normalization tool, Bioinformatics 38 (2022) 4449–4451.

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

[3] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2022.

[4] L. Lange, H. Adel, J. Strötgen, D. Klakow, CLIN-X: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain, Bioinformatics 38 (2022) 3267–3274.

[5] M. M. Agüero-Torales, C. Rodríguez Abellán, M. Carcajona Mata, J. I. Díaz Hernández, M. Solís López, A. Miranda-Escalada, S. López-Alvárez, J. Mira Prats, C. Castaño Moraga, D. Vilares, L. Chiruzzo, Overview of GenoVarDis at IberLEF 2024: NER of Genomic Variants and Related Diseases in Spanish, Procesamiento del Lenguaje Natural 73 (2024).

[6] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, 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, 2024.

[7] M. Villegas, A. Intxaurrondo, A. Gonzalez-Agirre, M. Marimon, M. Krallinger, The mespen resource for english-spanish medical machine translation and terminologies : Census of parallel corpora , glossaries and term translations, 2018. URL: https://api.semanticscholar.org/CorpusID:195876638.

[8] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, New York, NY, USA, 2019.