Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing

Matías Rojas[1,3], **Jocelyn Dunstan**[2,3,4], **and Fabián Villena**[1,3]
[1]Department of Computer Sciences, University of Chile.
[2]Initiative for Data & Artificial Intelligence, University of Chile.
[3]Center for Mathematical Modeling - CNRS IRL 2807, University of Chile.
[4]Millenium Institute for Intelligent Healthcare Engineering, ANID, Chile.
matias.rojas.g@ug.uchile.cl
{jdunstan, fabian.villena}@uchile.cl

Abstract

Word embeddings have been widely used in Natural Language Processing (NLP) tasks. Although these representations can capture the semantic information of words, they cannot learn the sequence-level semantics. This problem can be handled using contextual word embeddings derived from pre-trained language models, which have contributed to significant improvements in several NLP tasks. Further improvements are achieved when pre-training these models on domain-specific corpora. In this paper, we introduce Clinical Flair, a domain-specific language model trained on Spanish clinical narratives. To validate the quality of the contextual representations retrieved from our model, we tested them on four named entity recognition datasets belonging to the clinical and biomedical domains. Our experiments confirm that incorporating domain-specific embeddings into classical sequence labeling architectures improves model performance dramatically compared to general-domain embeddings, demonstrating the importance of having these resources available.

1 Introduction

Word embeddings are dense, semantically meaningful vector representations of a word. This method has proven to be a fundamental building block when constructing neural network-based architectures. However, the main drawback of using these embeddings is that they provide only a single representation of a given word across many documents. This is not optimal in practice, as the representation depends on the sentence in which the word appears. Contextual word embeddings address this problem by capturing syntactic and semantic information at the sentence level to represent words according to their context.

Contextualized embeddings are commonly retrieved from language models trained on giant text corpora. These models are usually composed of sequential or attention neural networks, which allows obtaining sentence-level semantics. This method has contributed to major advances in several NLP tasks such as named entity recognition, text classification, and relation extraction. Classic examples of contextual representation models are Flair (Akbik et al., 2018), ELMo (Peters et al., 2018), and BERT (Devlin et al., 2019).

Regarding specific domains such as clinical and biomedical, there are widely used models for the English language, such as BioBERT (Lee et al., 2020), BioELMo (Jin et al., 2019), and the PubMed version of Flair. These studies have shown that incorporating domain-specific contextual word embeddings contributes to a significant improvement in the performance of the models. However, although unstructured clinical texts are abundant in Spanish, there is still a significant lack of language models. Most of the domain-specific contextual representation models available for Spanish focus on data obtained from scientific articles and not from texts written in a more realistic context.

To fill this gap, we trained and publicly released Clinical Flair[1], a character-level language model trained on a corpus with real diagnoses in Spanish. To measure the potential impact of using these representations, we provide an empirical study of the effects of using language models trained on domain-specific against general-domain corpora. We evaluated the effectiveness of the proposed embeddings on four named entity recognition datasets belonging to the clinical and biomedical domain in Spanish. The results suggest that the embeddings obtained from our model contribute to achieving a better model performance compared to the general-domain contextualized embeddings by a wide margin.

[1] https://github.com/plncmm/spanish-clinical-flair

July 14, 2022 ©2022 Association for Computational Linguistics

2 Related Work

Language models allow us to generate high-quality representations of words based on their surrounding context, better known as contextual word embeddings. These models are usually trained with large corpora, either general-domain or domain-specific. Most of the available models have been trained with English resources, where the most popular ones are BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), GPT-2 (Radford et al., 2019), and Flair (Akbik et al., 2018).

As pointed out in Lee et al. (2020), building domain-specific language models allows to improve models performance compared to general-domain language models. In relation to biomedical information retrieval (IR) tasks in English, the most well-known architectures are BioBERT (Lee et al., 2020), Clinical BERT (Alsentzer et al., 2019), SciBERT (Beltagy et al., 2019), Pubmed BERT (Gu et al., 2022), BioELMo (Jin et al., 2019) and Pubmed Flair.

Regarding the clinical domain in Spanish, we found the models Biomedical Roberta (Carrino et al., 2022) and SciELO Flair (Akhtyamova et al., 2020). In the first case, the main difference with our model is that Biomedical Roberta was trained on a corpus formed by several biomedical and clinical corpora, while we only used clinical narratives. In the case of SciELO Flair, a point of differentiation is that they used data obtained from medical publications, whereas our data comes from primary care diagnoses. Moreover, they only tested their model on the PharmaCoNER corpus, created from the same data source they trained SciELO Flair. In contrast, we tested the effectiveness of our model using four clinical and biomedical datasets.

3 Methods

This section describes the clinical dataset used to train our language model, the details of the training process, and, finally, the task and datasets used in our experiments.

3.1 Clinical Flair

Flair (Akbik et al., 2018) is a character-level language model, which represents words as sequences of characters contextualized by the surrounded text. Flair authors created a method to obtain contextualized representations by retrieving the internal states of a bidirectional character-level LSTM. Specifically, the embedding is created by concatenating the output of the hidden state after the last character and before the first character of the word. This process allows obtaining the word context in the sentence in both directions.

We decided to use Flair instead of BERT because the character-level language model is beneficial for handling misspelled and out-of-vocabulary words, which are abundant in clinical and biomedical texts. This is because BERT is limited to a predefined vocabulary used to perform the tokenization. When a word is outside the vocabulary, the BERT model combines the embeddings of its subwords to compute the final representation, which may decrease the quality of the embeddings. This does not occur in the case of Flair, where each word has an embedding independent of its subword embeddings.

To create our clinical version of Flair, we used as a starting point the existing language models *es-forward* and *es-backward*. These models trained on a large corpus obtained from the Spanish Wikipedia are freely available in the Flair framework (Akbik et al., 2019). To incorporate key information from the clinical context, we fine-tuned these models on the Chilean Waiting List corpus (Báez et al., 2020), which is a clinical corpus created from real diagnoses from the Chilean public healthcare system.

The Chilean Waiting List corpus consists of $5,157,902$ free-text diagnostic suspicions comprising $14,057,401$ sentences and $68,541,727$ tokens. Although the general purpose of this dataset was to be a new resource for named entity recognition, it has also been used to obtain static word embeddings from the clinical domain (Villena et al., 2021b). These representations have boosted the model's performance in several clinical NLP tasks such as tumor encoding (Villena et al., 2021a) and named entity recognition (Báez et al., 2022).

We did not perform any pre-processing of the data for training our language model. The corpus was divided into 60% for training, 20% for validation, and 20% for testing. According to the suggestions of Flair authors, we set the maximum sentence length to 250, the mini-batches to 100 sentences, the maximum training epochs to $1,000$, and the learning rate to 20. The experiments were performed with a Tesla V100 GPU and 192 GB RAM. After one week of training, we reached a final perplexity value of 1.61 and 1.63 for our *es-clinical-forward* and *es-clinical-backward* models, respectively.

	CANTEMIST			PharmaCoNER			Clinical Trials			NUBes		
	Train	Test	Dev	Train	Test	Dev	Train	Test	Dev	Train	Test	Dev
Tokens	442,097	240,326	396,457	210,778	104,201	100,147	208,188	68,994	69,319	255,897	51,233	35,416
Sentences	19,397	11,168	18,165	8,177	3,976	3,790	12,555	4,506	4,550	13,802	2,762	1,840
Avg sentence length	22.8	21.5	21.8	25.8	26.2	26.4	16.6	15.3	15.3	18.5	18.6	19,2
Entities	6,347	3,596	5,948	3,821	1,876	1,926	24,224	7,717	8,258	17,122	3,548	2,293
Avg entity length	2.4	2.3	2.3	1.4	1.4	1.4	2.0	2.0	2.0	2.6	2.6	2.6

Table 1: Statistics of the NER datasets used in our experiments.

3.2 Datasets

To evaluate the quality of our contextual representations, we used the Named Entity Recognition (NER) task, which seeks to identify spans of text expressing references to predefined categories. Specifically, we performed our experiments on four NER corpora belonging to the clinical and biomedical domains. The statistics for each corpus are shown in Table 1.

- **CANTEMIST**[2] (**Miranda-Escalada et al., 2020**): An open annotated corpus that comprises 1,301 oncologic clinical case reports written in Spanish and manually annotated by clinical experts with mentions of tumor morphology. It contains a total of 48,730 sentences and 15,891 entity mentions.

- **PharmaCoNER**[3] (**Gonzalez-Agirre et al., 2019**): Biomedical corpus created for recognizing chemical and protein entities. It consists of 1,000 clinical cases with 7,623 entity mentions, corresponding to four entity types.

- **Clinical Trials**[4] (**Campillos-Llanos et al., 2021**): It consists of 1,200 texts collected from 500 abstracts of journal articles about clinical trials and 700 announcements of trial protocols. It comprises a total of 40,199 entity mentions, which belong to a subset of semantic groups from the Unified Medical Language System (UMLS).

- **NUBes**[5] (**Lima Lopez et al., 2020**): Biomedical corpus obtained from anonymized health records annotated with negation and uncertainty. It consists of 18,404 sentences, including 22,963 mentions of negation and uncertainty.

Parameter	Value
max epochs	150
optimizer	SGD
batch size	32
initial learning rate	0.1
word dropout	0.05
BiLSTM layers	1
BiLSTM hidden size	256

Table 2: Hyperparameters used in our experiments.

3.3 NER Model

To solve the NER task, we used the LSTM-CRF approach proposed by Lample et al. (2016), which is one of the most widely used architectures for sequence labeling tasks. The model consists of three main modules: the embedding layer, the encoding layer with a BiLSTM, and the classification layer, where the most likely sequence of labels is obtained using the CRF algorithm. Our contextualized embeddings were incorporated in the first layer, replacing traditional representations such as word and character-level embeddings.

To compare the performance of our language model, we used two baselines: the Spanish Flair model trained on the general domain using Wikipedia articles and the SciELO Flair model, which was trained over a subset of SciELO text.

In addition, it is worth mentioning that some of the datasets had nested entities, i.e., entities contained within other entity mentions (Finkel and Manning, 2009). Since traditional sequence labeling architectures cannot address this problem, we followed the simplifications made in previous work, keeping only the outermost entities in each nesting.

3.4 Settings

To select the best hyperparameters, we performed the random search strategy, which selects the best values by exhaustively testing different combinations of hyperparameters over a range of values. We measured the performance using the validation partition to establish the best combination.

[2] https://zenodo.org/record/3978041

[3] https://zenodo.org/record/4270158

[4] http://www.lllf.uam.es/ESP/nlpmedterm_en

[5] https://github.com/Vicomtech/NUBes-negation-uncertainty-biomedical-corpus

	Spanish Flair			SciELO Flair			Clinical Flair		
Dataset	P	R	F1	P	R	F1	P	R	F1
CANTEMIST	0.827 (0.002)	0.842 (0.003)	0.834 (0.001)	0.850 (0.001)	0.864 (0.001)	0.857 (0.001)	0.857 (0.004)	0.867 (0.001)	**0.862 (0.002)**
PharmaCoNER	0.876 (0.002)	0.849 (0.001)	0.862 (0.001)	0.905 (0.001)	0.889 (0.002)	**0.897 (0.001)**	0.901 (0.001)	0.875 (0.002)	0.888 (0.001)
Clinical Trials	0.809 (0.003)	0.815 (0.001)	0.812 (0.001)	0.814 (0.005)	0.832 (0.001)	0.823 (0.002)	0.836 (0.002)	0.834 (0.003)	**0.835 (0.001)**
NUBes	0.887 (0.002)	0.901 (0.003)	0.894 (0.001)	0.888 (0.002)	0.905 (0.001)	0.896 (0.001)	0.905 (0.002)	0.897 (0.001)	**0.901 (0.001)**

Table 3: Overall results on four clinical and biomedical NER datasets. Data shown are mean (SD).

In Table 2, we list the main hyperparameters used throughout our experiments, which were the ones that gave us the best results in most of the datasets. We trained the NER models using the SGD optimizer to a maximum of 150 epochs, with mini-batches of size 32 and a learning rate of 0.1. To control overfitting, we used the early stopping strategy and a dropout regularization of 0.05 after the embedding layer.

Performance was evaluated using precision, recall, and micro F1-score, which is the standard metric used in NER. This metric is strict since an entity is considered correct when both entity types and boundaries are predicted correctly. Three rounds of evaluation were computed using different seeds, reporting the mean and standard deviation. All the experiments were performed using the Flair framework, and the source code is available to reproduce our experiments[6].

4 Results

Table 3 shows the overall performance of the NER model comparing contextualized embeddings retrieved from our Clinical Flair model, Spanish Flair, and SciELO Flair. We can see that across all datasets, the performance of our model is superior to the model trained on a general domain, demonstrating the importance of incorporating contextualized embeddings trained on domain-specific corpora.

On the other hand, although we did not train our model on biomedical corpora, we observe that it is also beneficial for solving NER on those datasets. Although we did not outperform the SciELO Flair model in PharmaCoNER, we obtained competitive results. However, as mentioned in their paper, they selected a subset of SciELO texts to train the language model in line with the PharmaCoNER corpus. Therefore, we expected that their results would be superior.

Compared with Spanish Flair, the major difference occurs in CANTEMIST, reaching an average difference of +0.028, while the slightest difference is observed in NUBes with +0.007 according to the F1 measure. One possible reason for the similar performance between our model and Spanish Flair in NUBes is that, although the dataset belongs to the biomedical domain, the task aims to identify entities associated with negations and uncertainties; therefore, the target labels are general-domain and distant from the original corpus on which we trained our model.

Finally, and as expected, in both corpora belonging to the clinical domain CANTEMIST and Clinical Trials, our model outperforms both Spanish Flair and SciELO Flair. In the case of Clinical Trials, we reached an average difference of +0.023 and +0.012 compared to both models, respectively, while in the case of CANTEMIST, we obtained improvements of +0.028 and +0.005 according to the F1 measure.

5 Conclusions and Future Work

Despite the growing interest of the NLP research community in contextualized embeddings, there is still a lack of language models for the Spanish language, a gap that increases even more concerning domain-specific texts. To address this issue, this paper introduced Clinical Flair, a character-level language model for clinical NLP in Spanish. Specifically, we used a general-domain language model as a starting point and then fine-tuned it on Chilean clinical narratives. Our experimental results on four clinical and biomedical NER datasets show that incorporating our domain-specific embeddings outperforms by a wide margin the results obtained with general-domain embeddings, demonstrating the importance of having these resources available for languages not as widely explored.

Future work includes extending our study to other NLP tasks and using different combinations of embeddings, such as concatenating Word2vec or character-level embeddings. In addition, to provide a variety of contextual representation models for clinical texts, we are training a clinical version of BERT in Spanish. Although preliminary

[6] https://github.com/plncmm/clinical-flair

results have been inferior to those obtained with our Clinical Flair model, we expect to collect a larger clinical corpus to improve performance.

Acknowledgements

This work was funded by ANID Chile: Basal Funds for Center of Excellence FB210005 (CMM), Millennium Science Initiative Program ICN2021_004 (iHealth), and Fondecyt grant 11201250. This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02) and the Patagón supercomputer of Universidad Austral de Chile (FONDEQUIP EQM180042). We also acknowledge the help received from Kinan Martin and the reviewers.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Liliya Akhtyamova, Paloma Martínez, Karin Verspoor, and John Cardiff. 2020. Testing contextualized word embeddings to improve ner in spanish clinical case narratives. *IEEE Access*, 8:164717–164726.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Pablo Báez, Felipe Bravo-Marquez, Jocelyn Dunstan, Matías Rojas, and Fabián Villena. 2022. Automatic extraction of nested entities in clinical referrals in spanish. *ACM Trans. Comput. Healthcare*, 3(3).

Pablo Báez, Fabián Villena, Matías Rojas, Manuel Durán, and Jocelyn Dunstan. 2020. The Chilean waiting list corpus: a new resource for clinical named entity recognition in Spanish. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 291–300, Online. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Leonardo Campillos-Llanos, Ana Valverde-Mateos, Adrián Capllonch-Carrión, and Antonio Moreno-Sandoval. 2021. A clinical trials corpus annotated with umls entities to enhance the access to evidence-based medicine. *BMC Medical Informatics and Decision Making*, 21.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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. Association for Computational Linguistics.

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 141–150, Singapore. Association for Computational Linguistics.

Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. 2019. PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 82–89, Minneapolis, USA. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016.

Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Salvador Lima Lopez, Naiara Perez, Montse Cuadros, and German Rigau. 2020. NUBes: A corpus of negation and uncertainty in Spanish clinical texts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5772–5781, Marseille, France. European Language Resources Association.

A Miranda-Escalada, 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*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Fabián Villena, Pablo Báez, Sergio Peñafiel, Matías Rojas, Inti Paredes, and Jocelyn Dunstan. 2021a. Automatic support system for tumor coding in pathology reports in spanish. *SSRN Electronic Journal*.

Fabian Villena, Jorge Perez, Rene Lagos, and Jocelyn Dunstan. 2021b. Supporting the classification of patients in public hospitals in chile by designing, deploying and validating a system based on natural language processing. *BMC Medical Informatics and Decision Making*, 21(1):1–11.