Development of pre-trained language models for clinical NLP in Spanish

Claudio Aracena[1,2] **and Jocelyn Dunstan**[2,3,4,5]
[1]Faculty of Physical and Mathematical Sciences, University of Chile
[2]Millennium Institute Foundational Research on Data, Chile
[3]Department of Computer Science, Pontifical Catholic University of Chile
[4]Institute for Computational Mathematics, Pontifical Catholic University of Chile
[5]Center for Mathematical Modeling, University of Chile
claudio.aracena@uchile.cl, jdunstan@uc.cl

Abstract

Clinical natural language processing aims to tackle language and prediction tasks using text from medical practice, such as clinical notes, prescriptions, and discharge summaries. Several approaches have been tried to deal with these tasks. Since 2017, pre-trained language models (PLMs) have achieved state-of-the-art performance in many tasks. However, most works have been developed in English. This PhD research proposal addresses the development of PLMs for clinical NLP in Spanish. To carry out this study, we will build a clinical corpus big enough to implement a functional PLM. We will test several PLM architectures and evaluate them with language and prediction tasks. The novelty of this work lies in the use of only clinical text, while previous clinical PLMs have used a mix of general, biomedical, and clinical text.

1 Introduction

Clinical text is one of the richest forms of information in electronic health records. Therefore, developing tools to extract useful information from clinical text has become relevant in clinical natural language processing (NLP). However, processing unstructured text is challenging due to the complexity of human languages. Moreover, the clinical text has its own complexities, including non-standard abbreviations, misspellings, specific vocabulary, and jargon (Dalianis, 2018).

Clinical NLP aims to address several tasks in this complex scenario. These tasks can range from language tasks such as extracting entities, text classification, and relation extraction, among others, to prediction tasks such as predicting patient mortality, length of hospital stay, unplanned readmissions, etc. Several works have been carried out to tackle these tasks generating specific models.

However, since 2017, the NLP field has worked towards the creation of pre-trained language models (PLMs) that can be fine-tuned for any specific downstream task. These language models are built for a much simpler task, such as next-word or masked-word prediction in a huge amount of text. This process, known as pre-training, allows the language model to acquire language understanding that can be used for any text-related task (Tunstall et al., 2022).

As soon as the NLP field started to work in PLMs, clinical NLP introduced this type of model into its set of techniques to improve performance in its own tasks. Some examples of clinical PLMs are two different versions of ClinicalBERT (Alsentzer et al., 2019; Huang et al., 2020). These models show a significant improvement in language tasks and a moderate improvement in prediction tasks.

Most of the research in clinical NLP has been done for text written in English, but not so much for other languages (Névéol et al., 2018). In Spanish, some publicly available PLMs relevant to clinical NLP are bsc-bio-ehr-es (Carrino et al., 2022) and Spanish Clinical Flair (Rojas et al., 2022). These PLMs were pre-trained heavily in general and biomedical text with some additions of clinical text. Despite this drawback, they outperform general and biomedical PLMs in language tasks.

There are two approaches to evaluate PLMs, intrinsic and extrinsic. An intrinsic approach measures the model quality independently of an application in any specific task. Examples of intrinsic metrics are perplexity and word similarity. Meanwhile, an extrinsic approach evaluates the model performance in downstream tasks, such as text classification or named entity recognition (NER). Extrinsic evaluations are more expensive and time-consuming than intrinsic evaluations (Jurafsky and Martin, 2021).

In Biomedical NLP, some relevant benchmarks have been developed for extrinsic evaluations, such as BLUE (Peng et al., 2019) and BLURB (Gu et al., 2021). Given their close relation to the clinical context, they are also used to evaluate clinical PLMs.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 52–60

However, these benchmarks are built with tasks in English. There are no such benchmarks in Spanish, thus the research community evaluates their models in downstream tasks for biomedical and clinical corpora. Most of these downstream tasks are language tasks, and few works focus on prediction tasks.

In this research proposal, we aim to develop clinical PLMs in Spanish using mostly clinical text and evaluate them with intrinsic and extrinsic approaches. The expected results are as follows:

- The development of a PLM in Spanish with a large clinical corpus, instead of general or biomedical text, will improve performance in intrinsic and extrinsic clinical evaluations.
- The clinical PLM will perform similarly in language tasks compared to existing PLMs, but it will outperform them for prediction tasks. This work focuses on prediction tasks due to their applicability to real problems. Thus, the expected results aim for better performance in prediction tasks.

This research proposal shows a literature review of PLMs architectures for general, biomedical, and clinical domains. Then, it states the research questions and their implications. Later, it describes the methodology to carry out this research and details the expected results.

2 Literature review

Since the creation of transformers (Vaswani et al., 2017) and ULMFiT (Howard and Ruder, 2018), we have witnessed the development of several PLMs that have become state-of-the-art on different tasks. Starting with GPT (Radford et al., 2018) and BERT (Devlin et al., 2019), continuing with RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021b,a), among others, the research community has focused on finding new architectures that can beat current benchmarks and apply them in many areas as possible.

One of the areas where it is possible to find several PLMs is health. Just a few months after the publication of BERT, PLMs using clinical text were released (Alsentzer et al., 2019; Huang et al., 2020). Moreover, even though PLMs are built mostly for English text, it is still possible to find clinical and biomedical PLMs for Spanish (Carrino et al., 2022; Rojas et al., 2022).

This section introduces PLMs in the clinical and biomedical fields and describes significant clinical and biomedical PLMs in English and Spanish. Also, it shows some evaluation benchmarks and tasks for biomedical and clinical NLP in English and Spanish.

2.1 Biomedical and Clinical PLMs for English

Biomedical PLMs refer to models pre-trained in medical text from academic sources, such as scientific publications. Meanwhile, clinical PLMs refer to models pre-trained in medical text from the medical practice, such as clinical notes and prescriptions.

2.1.1 BioBERT

Following BERT's success, other fields created their own versions of this model. BioBERT was the first BERT-based PLM pre-trained in a biomedical corpus (Lee et al., 2020). Using BERT architecture, BioBERT was pre-trained with the English Wikipedia, BookCorpus (Zhu et al., 2015), PubMed abstracts, and PubMedCentral (PMC) full-text articles, totaling 21 billion words. BioBERT was fine-tuned on a series of biomedical NLP tasks, such as NER, relation extraction (RE), and QA. BioBERT achieved state-of-the-art in most of the tasks under study and significantly outperforms BERT, showing that pre-training in specific domain data is a key step for downstream tasks' performance.

2.1.2 ClinicalBERT (2019)

One of the first BERT-based clinical PLMs was ClinicalBERT (Alsentzer et al., 2019). Unlike BioBERT, ClinicalBERT used clinical text such as physician notes and discharge summaries, both extracted from MIMIC-III database (Johnson et al., 2016). Several versions of ClinicalBERT were implemented using a technique currently known as continual pre-training. Continual pre-training means starting from an existing PLM to continue the pre-training process on additional data. In the case of ClinicalBERT, BERT and BioBERT were used as starting points, and clinical notes and discharge summaries as additional data to continue pre-training. One task of NLI and four NER tasks were used for the fine-tuning process. In just two tasks, Bio+ClinicalBERT (ClinicalBERT starting from BioBERT) outperforms BioBERT, showing that adding additional clinical text in continual pre-training can help performance, but not in all

cases.in all cases.

2.1.3 ClinicalBERT (2020)

Concurrently with the previous work, another ClinicalBERT was being developed (Huang et al., 2020). In this case, a continual pre-training process was implemented using BERT as starting point and clinical notes from MIMIC-III as additional data, similar to ClinicalBERT (2019). However, a readmission task was used for the fine-tuning process. This task aims to predict if a patient will be readmitted in the next 30 days after discharge, a clinical-specific task. For readmission prediction, the ClinicalBERT (2020) output for the classification token [CLS] is passed to a classification layer with a sigmoid function. ClinicalBERT (2020) outperforms BERT and the other two methods by more than 2% in AUC.

2.1.4 PubMedBERT

One of the latest and state-of-the-art biomedical PLMs is PubMedBERT (Gu et al., 2021). The assumption that general domain text can help pre-training to introduce general language knowledge into PLMs is challenged by this work. In mixed-domain pre-training, the vocabulary and the pre-training corpus come from general-domain text. Some models, like BioBERT, add text from biomedical sources to the pre-training corpus, and others, like ClinicalBERT, do continual pre-training over clinical text. On the other hand, domain-specific pre-training from scratch generates both the vocabulary and the pre-training corpus from specific-domain text. PubMedBERT implements the latter approach.

The domain-specific pre-training approach requires a large amount of text, which in the field of biomedicine can be found in PubMed. PubMedBERT pre-training corpus consists of 14 million abstracts, and 3.2 billion words, totaling 21 GB of uncompressed text. Out of 13 tasks, including NER, RE, QA, sentence similarity (SS), and document classification, PubMedBERT outperforms 11, showing that domain-specific pre-training is a better option for downstream tasks' performance.

2.2 General-domain, Biomedical, and Clinical PLMs for Spanish

2.2.1 BETO

The first implementation of BERT for Spanish is called BETO (Cañete et al., 2020). BETO used the same architecture that the original BERT model, but for the pre-training process it used some changes introduced by RoBERTa such as dynamic masking. Spanish Wikipedia and the Spanish parts of the OPUS Project (Tiedemann, 2012) were used as a pre-training corpus, totaling 3 billion words. A benchmark of tasks, GLUES (Cañete et al., 2020), was built to compare BETO to a multilingual implementation of BERT, mBERT (Wu and Dredze, 2019). BETO outperforms most of the mBERT results, excluding some QA tasks. These results show that pre-training with Spanish text improves performance for most downstream tasks compared with multilingual PLM.

2.2.2 RoBERTa-bne

MarIA is a family of PLMs for Spanish (Gutiérrez-Fandiño et al., 2022). One PLM of interest in MarIA is the implementation of RoBERTa with Spanish text, RoBERTa-bne. RoBERTa-bne was pre-trained with text extracted from The National Library of Spain (Biblioteca Nacional de España or BNE). The pre-trained corpus consists of more than 200 million documents, and 135 billion tokens, totaling 570GB of uncompressed text. Two versions of RoBERTa-bne. Out of nine tasks, RoBERTa-bne outperforms eight compared to BETO and mBERT.

2.2.3 Bsc-bio-es and bsc-bio-ehr-es

The previous PLMs for Spanish were pre-trained in general-domain text. bsc-bio-es and bsc-bio-ehr-es are the first PLMs trained with exclusively biomedical and clinical text in Spanish (Carrino et al., 2022). Two corpora were built for this purpose, biomedical and clinical. The biomedical corpus consists of 2.5 million documents and 1.1 billion tokens, and the clinical corpus consists of 514k documents and 95 million tokens. Bsc-bio-es was pre-trained only with the biomedical corpus and bsc-bio-ehr-es with the biomedical and clinical corpora. The reason behind this design decision is two-fold; the clinical corpus is too small to create a functional PLM by itself, and to assess if adding a small clinical corpus to a large biomedical corpus has a positive impact on clinical NLP tasks.

Three tasks were used to benchmark the PLMs against others such as BETO-Galén, mBERT, mBERT-Galén, RoBERTa-bne, among others. BETO-Galén and mBERT-Galén are BETO and mBERT versions with continual pre-training in the Galén Oncology corpus (López-García et al., 2021), respectively. bsc-bio-ehr-es outperforms all other PLMs, including bsc-bio-es, in two tasks, and bsc-bio-es came in second place. For the remaining

task, another PLM, XLM-R-Galén (a continual pre-training version of XML-R (Conneau et al., 2020), a multilingual version of RoBERTa) outperforms all the other PLMs, but bsc-bio-es and bsc-bio-ehr-es came in second and third place, respectively.

These results show that using only biomedical and clinical corpora for the pre-training process improves performance compared to general-domain text in Spanish. This result is congruent with PubMedBERT findings for English. The results remain true even when the comparison is made with RoBERTa-bne, a model pre-trained with 100 times more text than bsc-bio-es and bsc-bio-ehr-es. Interestingly, XLM-R-Galén, trained in more than 2TB of general-domain data in 100 languages, has the best results for one task, but bsc-bio-es and bsc-bio-ehr-es are around 0.1% of F-score away.

2.2.4 Clinical Flair

Every PLM shown until this point has its architecture built upon transformers (Vaswani et al., 2017). However, there are PLM based on other architectures, such as Flair. Flair is a character-level language model, representing words as sequences of characters contextualized by close words. Flair uses the internal states of a bidirectional character-level LSTM to obtain contextualized word representations (Akbik et al., 2018).

Clinical Flair for Spanish (Rojas et al., 2022) is a continual pre-training version of a Flair implementation in Spanish (Akbik et al., 2019), which was trained over the Spanish Wikipedia. As additional data for pre-training, the Chilean Waiting List corpus was used (Báez et al., 2020; Báez et al., 2022). The corpus consists of 5 million diagnostic suspicions and 68 million words. Four NER tasks were used to evaluate the Clinical Flair. As Flair generates contextualized embeddings, an algorithm has to be used to solve NER tasks. The LSTM-CRF technique was used (Lample et al., 2016). Clinical Flair outperforms other Spanish Flair models in three tasks. For the remaining task, SciELO Flair (Akhtyamova et al., 2020), a biomedical Flair PLM, had the best results. These results also show the importance of domain-specific data, even when using continual pre-training.

2.3 Biomedical and Clinical PLMs evaluation

As mentioned in section 1, there are two approaches to evaluate PLMs, intrinsic and extrinsic. The intrinsic approach measures PLM representation's quality independently of any downstream task. Meanwhile, the extrinsic approach evaluates the PLM quality using downstream tasks.

2.3.1 Intrinsic evaluation

One of the most used intrinsic metrics is perplexity, which is the inverse probability of a word sequence normalized by the number of words (Jurafsky and Martin, 2021). Intuitively, the perplexity measures how well a language model predicts a sequence of words. Another metric is a medical graph-based intrinsic test, specifically built for biomedical and clinical PLMs. This intrinsic test proposes that using semantic distances between medical concepts extracted from a medical knowledge graph can help to measure the quality of representations made by the PLMs (Aracena et al., 2022). Both metrics can be applied to English and Spanish languages.

2.3.2 Extrinsic evaluation

In terms of extrinsic metrics, there are several benchmarks, such as the biomedical language understanding evaluation benchmark (BLUE) (Peng et al., 2019) and the biomedical language understanding reasoning benchmark (BLURB) (Gu et al., 2021). However, these benchmarks are based on datasets in English. In addition to the mentioned benchmarks, BigBio (Fries et al., 2022), a framework that gathers 126+ biomedical NLP datasets, covering 13 task categories and 10+ languages, can be used for extrinsic evaluation.

Unlike previous benchmarks, no biomedical nor clinical benchmark has been built with Spanish texts. However, there are NER tasks commonly used to evaluate PLMs built with Spanish corpora. Following is a list of tasks of interest for this proposal, and in Table 1, their details.:

CANTEMIST-NER[1] (Miranda-Escalada et al., 2020): annotated corpus with tumor morphology mentions in 1,301 oncological clinical case reports.

PharmaCoNER[2] (Gonzalez-Agirre et al., 2019): annotated corpus with entities such as chemical compounds and drugs in 1,000 clinical case studies.

CT-EBM-SP[3] (Campillos-Llanos et al., 2021): annotated corpus with UMLS entities in 1,200 texts about clinical trials studies and announcements.

The Chilean Waiting List Corpus[4] (Báez et al., 2020; Báez et al., 2022): annotated corpus with

[1] https://zenodo.org/record/3978041
[2] https://zenodo.org/record/4270158
[3] http://www.lllf.uam.es/ESP/nlpmedterm$_en$
[4] https://zenodo.org/record/5591011

	CANTEMIST-NER			PharmaCoNER			CT-EBM-SP		
	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
Sentences	19,397	11,168	18,165	8,177	3,976	3,790	12,555	4,506	4,550
Avg sentence length	22.8	21.5	21.8	25.8	26.2	26.4	16.6	15.3	15.3
Entities	6,347	3,596	5,948	3,821	1,876	1,926	24,224	7,717	8,258
Avg entity length	2.4	2.3	2.3	1.4	1.4	1.4	2.0	2.0	2.0

	Chilean Waiting List			NUBes		
	Train	Test	Dev	Train	Test	Dev
Tokens	291,561	36,963	34,987	255,897	51,233	35,416
Sentences	15,290	1,912	1,911	13,802	2,762	1,840
Avg sentence length	19.07	19.33	18.31	18.5	18.6	19,2
Entities	69,847	8,837	8,340	17,122	3,548	2,293
Avg entity length	2.7	2.7	2.7	2.6	2.6	2.6

Table 1: Statistics of the NER tasks.

clinical entities such as findings, procedures, medications, etc. in 10,000 anonymized referrals for specialty consultations from the waiting list in Chilean public hospitals.

NUBes[5] (Lima Lopez et al., 2020): annotated corpus with negation and uncertainty entities in anonymized health records.

3 Research questions

From the literature review, it is possible to identify some research opportunities. These opportunities are mostly related to which data we can use to pre-train PLMs and which architecture best fits a specific task. Also, the evaluation process should be clear and easy for comparison, which in the case of clinical NLP in Spanish, there are still missing parts. Additionally, there is an opportunity to advance clinical NLP in Spanish by adopting strategies used in English and joining current efforts in the field.

From the identified opportunities, we can state the research questions as follows:

1. Will it be better to pre-train a PLM exclusively with clinical data compared to combinations of clinical and biomedical data for solving downstream clinical tasks in Spanish?

2. Is there a specific PLM architecture that outperforms others for solving downstream clinical tasks in Spanish?

To answer question one, we have to build a clinical corpus in Spanish big enough to implement a functional PLM. However, building a clinical corpus is a difficult task since there are privacy and confidentiality issues that may arise. The PLM trained over this corpus will be compared to biomedical and clinical PLMs for Spanish described in the literature review.

To answer question two, we have to test several PLMs architectures over the generated corpus. The literature review shows some interesting architecture to try in this research. However, as PLM architectures are being developed periodically, there is always the risk of becoming obsolete.

For both questions, we have to create an evaluation process that can include clinical intrinsic and extrinsic downstream tasks, as well as, language and prediction tasks. This evaluation process will allow us to clarify the research questions, but probably the answers will depend on the several tasks that we will include.

Hopefully, answering these research questions will contribute to the current efforts made by other laboratories working in clinical NLP in Spanish and languages other than English.

4 Methods

The methodology for this research proposal consists of three parts: build a clinical corpus for pre-training PLMs in Spanish, test different PLM architectures, and evaluate the implemented PLMs. This section will explain in detail every part.

4.1 Build a clinical corpus in Spanish

The literature review shows several corpora that can be used to pre-train PLMs. However, clinical corpora are hard to find, and some publications do

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

not release them given privacy and confidentiality issues. For example, the clinical corpus of bsc-bio-ehr-es is not publicly available, even though the biomedical corpora are available to download.

For this research proposal, an agreement with two medical institutions is in place. The agreement allows us to access clinical notes from both institutions. As a safety measure, we cannot use the data outside each institution to avoid unexpected privacy issues. We expect to use both sources to build two clinical corpora, one for each institution. Depending on the data quality of each source we will have to adjust the pre-processing steps to build each corpus. According to the authors of bsc-bio-ehr-es, the clinical corpus was left in its original form. Therefore, no pre-processing will be used in the first version of a PLM. Nevertheless, it is foreseeable that quality issues can appear, thus a cleaning process will be carried out for the second version.

Preliminary analysis of the data of the clinical institution shows that clinical notes consist of nearly one billion words. This situation shows that functional PLMs could be built from scratch.

4.2 Test different PLM architectures

As seen in the literature review, transformer-based architectures, such as BERT and RoBERTa, are the most common in building biomedical and clinical PLMs. In this proposal, we will test BERT and RoBERTa architectures as baselines. Apart from those two, DeBERTa is a relevant architecture to test, since it introduces important changes to the BERT architecture and it outperforms BERT and RoBERTa for general-domain downstream tasks. Additionally, we will test a non-transformer-based architecture, Flair, as it has not been previously pre-trained in clinical data exclusively, which can improve performance in downstream tasks. After this part, eight clinical PLMs will be implemented. Those PLMs will be BERT, RoBERTa, DeBERTa, and Flair versions for each clinical corpus.

4.3 Evaluate implemented PLMs

The evaluation process of the implemented PLMs has to consider intrinsic and extrinsic tasks as well as language and prediction tasks. Intrinsic tasks allow us to measure the internal quality of PLM independently of downstream tasks, so it is a good way to compare transformer-based architectures with others. Extrinsic tasks allow measuring performance in relevant downstream tasks for clinical NLP. Some of these tasks can be language tasks such as NER, QA, and document classification, among others, which are important to build tools to extract information from clinical text. Other tasks are prediction tasks such as prediction of unplanned readmissions, treatment length, diagnostic, etc., which support the decision-making of medical personnel.

For intrinsic tasks, we will use perplexity and the medical graph-based intrinsic test. The medical graph-based intrinsic test will be used to compare PLM where it is expected to have better results for clinical and biomedical PLM compared to general ones. For extrinsic tasks, we will use NER tasks such as CANTEMIST-NER, PharmaCoNER, CT-EBM-SP, NUBes, and the Chilean Waiting List. Additionally, we will use prediction tasks such as predicting unplanned readmissions and treatment length.

The implemented PLMs will be compared with existing and available general, biomedical and clinical PLMs using the evaluation process. Those PLMs will be BETO, mBERT, RoBERTa-bne, XLM-R, bsc-bio-es, bsc-bio-ehr-es, and clinical Flair.

5 Expected results

The expected results for this proposal are as follows:

- The implemented clinical PLM, mentioned in subsection 4.2, will outperform existing general, biomedical, and clinical PLM, mentioned in subsection 4.3. This result is expected given that evaluation tasks are mostly clinical-related and the new corpora are clinical.

- The implemented clinical PLM will have similar performance in language tasks compared to existing biomedical and clinical PLM, but it will outperform them for prediction tasks. This result is expected given that language tasks measure the quality of PLM in linguistic tasks related to the clinical context, and most existing biomedical and clinical PLM were pre-trained in corpora with some clinical context. However, prediction tasks measure how well PLM can extract information to predict a clinical outcome, and it is reasonable to believe that a PLM pre-trained only with

a clinical corpus will outperform PLM pretrained with a combination of biomedical and clinical text.

- The implemented PLM with DeBERTa architecture will outperform the RoBERTa version, and the latter PLM will outperform the BERT version. Given the previous performance for general, biomedical, and clinical PLM for English and Spanish, these results are expected. There is not enough evidence to draw a similar conclusion for the implemented PLM with Flair architecture.

- As an indirect result, if an implemented PLM considerably improves the evaluation in the task of prediction of treatment length, it could be integrated into the current institutions' systems to replace the already-in-use machine learning models.

6 Conclusions

This work presents a PhD research proposal for developing clinical PLMs in Spanish that can outperform general-domain, biomedical, and current clinical PLMs. As a result, it is expected to answer the research questions regarding whether using only clinical text to create a PLM can outperform current approaches and if a particular PLM architecture is better for clinical purposes.

We expect this work can help to understand how PLMs work in the clinical domain and in languages other than English.

Limitations

Limitations of this work can be listed as follows:

- This work will not test larger PLM architectures, such as large versions of BERT, RoBERTa, and DeBERTa, given the availability of computational resources and data.

- This paper focuses on encoder-only PLM architectures. This type of architecture has been heavily studied for biomedical and clinical PLMs and we continue with that line of research. However, decoder or encoder-decoder architectures can be tried to solve the same tasks.

- The clinical corpora to be built can introduce non-standard abbreviations and jargon unique to the clinical institutions providing the data. This can lower the performance of language tasks in other contexts.

- The clinical text will not be released given confidentiality issues which increment the gap between advances in general-domain and clinical NLP research.

Ethics Statement

We state that this work complies with the ACL Code of Ethics. We believe this work could help the research community with new information about clinical and biomedical PLM. The data for this work will not be released and will be used according to the ethical and confidentiality standards provided by the clinical institutions.

Acknowledgements

This work was funded by ANID Chile: Basal Funds for Center of Excellence FB210005 (CMM); Millennium Science Initiative Program ICN17_002 (IMFD) and ICN2021_004 (iHealth), Fondecyt grant 11201250, and National Doctoral Scholarships 21211659 (Claudio Aracena).

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. Conference Name: IEEE Access.

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.

Claudio Aracena, Fabián Villena, Matías Rojas, and Jocelyn Dunstan. 2022. A knowledge-graph-based intrinsic test for benchmarking medical concept embeddings and pretrained language models. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis*, Abu Dhabi, UAE. 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 Transactions on Computing for Healthcare (HEALTH)*, 3(3):1–22.

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.

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(1):69.

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.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Ho Jou-Hui, Kang Hojin, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *Practical ML for Developing Countries Workshop @ ICLR 2020*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Hercules Dalianis. 2018. *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer International Publishing.

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.

Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Myungsun Kang, Ruisi Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose David Posada, John Michael Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Michio Broad, Yanis Labrak, Shlok S Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. 2022. BigBio: A framework for data-centric biomedical natural language processing. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):2:1–2:23.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68(0):39–60. Number: 0.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. arXiv.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035. Number: 1 Publisher: Nature Publishing Group.

Daniel Jurafsky and James H. Martin. 2021. *Speech and Language Processing*.

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 Twelfth Language Resources and Evaluation Conference*, pages 5772–5781, Marseille, France. European Language Resources Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Guillermo López-García, José M. Jerez, Nuria Ribelles, Emilio Alba, and Francisco J. Veredas. 2021. Transformers for Clinical Coding in Spanish. *IEEE Access*, 9:72387–72397. Conference Name: IEEE Access.

Antonio Miranda-Escalada, Eulàlia Farré, and Martin Krallinger. 2020. Cantemist corpus: gold standard of oncology clinical cases annotated with CIE-O 3 terminology. Version Number: 1.6 Type: dataset.

Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical Natural Language Processing in languages other than English: opportunities and challenges. *Journal of Biomedical Semantics*, 9(1):12.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Matías Rojas, Jocelyn Dunstan, and Fabián Villena. 2022. Clinical Flair: A Pre-Trained Language Model for Spanish Clinical Natural Language Processing. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 87–92, Seattle, WA. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers*. O'Reilly Media, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. pages 19–27.