Pre-trained Biomedical Language Models for Clinical NLP in Spanish

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño
Jordi Armengol-Estapé, **Joaquín Silveira-Ocampo**, **Alfonso Valencia**
Aitor Gonzalez-Agirre and **Marta Villegas**
Barcelona Supercomputing Center
marta.villegas@bsc.es

Abstract

This work presents the first large-scale biomedical Spanish language models trained from scratch, using large biomedical corpora consisting of a total of 1.1B tokens and an EHR corpus of 95M tokens. We compared them against general-domain and other domain-specific models for Spanish on three clinical NER tasks. As main results, our models are superior across the NER tasks, rendering them more convenient for clinical NLP applications. Furthermore, our findings indicate that when enough data is available, pre-training from scratch is better than continual pre-training when tested on clinical tasks, raising an exciting research question about which approach is optimal. Our models and fine-tuning scripts are publicly available at HuggingFace and GitHub.

1 Introduction and Background

The success of Transformer-based models in the general domain (Devlin et al., 2019) soon encouraged the development of language models for domain-specific scenarios (Chalkidis et al., 2020; Gutiérrez-Fandiño et al., 2021; Tai et al., 2020; Araci, 2019; Lee and Hsiang, 2019). Specifically, in the biomedical domain, there has been a proliferation of models (Peng et al., 2019; Beltagy et al., 2019; Alsentzer et al., 2019; Gu et al., 2021) since the first BioBERT (Lee et al., 2019) model was published. Unfortunately, there is still a significant lack of biomedical and clinical models in languages other than English, despite the increasing efforts of the NLP community (Névéol et al., 2014; Schneider et al., 2020). Consequently, general-domain pre-trained language models supporting Spanish, such as mBERT (Devlin et al., 2019) and BETO (Cañete et al., 2020), have been often used as a proxy to build domain-specific systems in the absence of genuine alternatives. For instance, Sun and Yang (2019) used mBERT and BioBERT on the PharmaCoNER (Gonzalez-Agirre et al., 2019) dataset, using a fine-tuning strategy aimed to maximize the results.

Very recently, new pre-trained clinical language models for Spanish have been published (López-García et al., 2021) by further pre-training the mBERT, BETO and XLM-RoBERTa (Conneau et al., 2020) models with a corpus of Spanish clinical cases with about 64M tokens. In our work, we go one step further to address the language gap for Spanish and train two Transformer-based language models from scratch. We employed biomedical and clinical corpora (including clinical texts) gathered by ourselves. We evaluated our models with three different Named Entity Recognition (NER) tasks, since NER constitutes a core task in many clinical NLP scenarios. They obtained significant gains over the general-domain models, and matched or outperformed the domain-specific models in all tasks.

2 Corpora

We built two corpora of very different sizes and nature: an Electronic Health Record (EHR) corpus and a biomedical one. The **EHR corpus** contains 95M tokens from more than 514k clinical documents (including discharge reports, clinical course notes and X-ray reports). The **biomedical corpus** includes Spanish data from a variety of sources for a total of 1.1B tokens across 2,5M documents, namely:

- **Medical crawler**:[1] Crawler of more than 3,000 URLs belonging to Spanish biomedical and health domains (Carrino et al., 2021).
- **Clinical cases misc.**: A miscellany of medical content, essentially clinical cases. Note that a clinical case report is different from a scientific publication where medical practitioners share patient cases and it is also different from a clinical note or document.

[1] https://zenodo.org/record/4561970

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 193–199

- **Scielo**:[2] Scientific publications written in Spanish crawled from the Spanish SciELO server in 2017.
- **BARR2 Background**:[3] Biomedical Abbreviation Recognition and Resolution (BARR2) containing Spanish clinical case study sections from a variety of clinical disciplines.
- **Wikipedia (Life Sciences)**: Wikipedia articles crawled on 04/01/2021 with the Wikipedia API python library[4] starting from the "Ciencias_de_la_vida" category up to a maximum of 5 subcategories. Multiple links to the same article are discarded to avoid repeated content.
- **Patents**: Google Patent in Medical Domain for Spain (Spanish). The accepted codes (Medical Domain) for JSON files of patents are: "A61B", "A61C","A61F", "A61H", "A61K", "A61L","A61M", and "A61P".
- **EMEA**:[5] Spanish-side documents extracted from parallel corpora made out of PDF documents from the European Medicines Agency.
- **Mespen (MedlinePlus)**:[6] Spanish-side articles extracted from a collection of Spanish-English parallel corpus consisting of biomedical scientific literature. The collection of parallel resources are aggregated from the MedlinePlus source.
- **PubMed**: Open-access Spanish abstracts from the PubMed repository crawled in 2017.

For each biomedical resource, we applied a cleaning pipeline with customized operations designed to read data in different formats, split it into sentences, detect the language, remove noisy and ill-formed sentences, deduplicate and eventually output the data with their original document boundaries. Finally, to remove repetitive content, we concatenated the entire corpus and deduplicated it again, obtaining about 1.1B words. These preprocessing steps were applied to all data except the EHR corpus, which was left in its original form. Table 1 shows detailed statistics of each component of the corpus.

[2] https://zenodo.org/record/2541681

[3] https://temu.bsc.es/BARR2/downloads/background_set.raw_text.tar.bz2

[4] https://github.com/martin-majlis/Wikipedia-API/

[5] http://opus.nlpl.eu/download.php?f=EMEA/v3/moses/en-es.txt.zip

[6] https://zenodo.org/record/3562536

Source	No. tokens
Medical crawler	903,558,136
Clinical cases misc.	102,855,267
EHRs documents*	95,267,204
Scielo	60,007,289
BARR2 Background	24,516,442
Wikipedia (Life Sciences)	13,890,501
Patents	13,463,387
EMEA	5,377,448
Mespen (MedlinePlus)	4,166,077
PubMed	1,858,966

Table 1: List of individual sources in the training corpora. The number of tokens refers to *white-spaced* tokens on cleaned untokenized text. Documents from the EHR corpus are marked with an asterisk.

3 Models Pre-training

The models presented in this work were pre-trained from scratch employing a RoBERTa (Liu et al., 2019) base model with 12 self-attention layers. Following the original training, we only used Masked Language Modeling (MLM) as the pre-training objective with Subword Masking (SWM), as in (Liu et al., 2019).

We tokenized the training corpus with the Byte-Level BPE algorithm (Radford et al., 2019), employed in the original RoBERTa, and learned a vocabulary of 50,262 tokens.

We run the training for 48 hours on 16 NVIDIA V100 GPUs of 16GB VRAM, using Adam optimizer (Kingma and Ba, 2015) with a peak learning rate of 0.0005, 10,000 warm-up steps and an effective batch size of 2,048 sentences.[7] Other hyper-parameters were left in their default values as in the original RoBERTa training configuration. Training was performed at the document level, preserving document boundaries.[8] We performed a train-validation split based on the number of documents, choosing a total of 2,000 documents for the validation set, corresponding to less than 1% of the entire corpus' documents. We then select the model with the lowest perplexity on the validation set as the best model.

We used the corpora described in the previous section to produce two RoBERTa models: a biomedical language model training

[7] Through gradient accumulation as implemented in Fairseq (Ott et al., 2019)

[8] We believe document-level training may be crucial to promote the modelling of long-range dependencies and push the model towards the comprehension of entire documents.

only with the so-called biomedical resources (bsc-bio-es),[9] and a BIO-EHR language model that uses both the biomedical and EHR corpus (bsc-bio-ehr-es).[10] We trained the latter model, the biomedical-EHR, to assess if adding a relatively small EHR data to a large-scale corpora has a positive impact on real-world clinical NLP tasks.

4 NER Fine-tuning

We tested and evaluated our models by fine-tuning the NER task, a key component of information extraction tasks in the clinical domain. Indeed, we used it as a testbed to evaluate the effectiveness of our pre-trained models. Following the usual fine-tuning method, employed both for general-domain models (Devlin et al., 2019; Liu et al., 2019) and domain-specific ones (Lee et al., 2019), we added a standard linear layer as a token classification head, and the BIO tagging schema (Sang and Buchholz, 2000) to solve the NER tasks. During fine-tuning, both the pre-trained model and the classification layer's parameters are learned with stochastic gradient descent. We used an Adam (Kingma and Ba, 2015) optimizer and searched for an optimal learning rate out of [8e-6, 1e-5, 2e-5, 3e-5, 5e-5] with linear decay and no warm-up steps. We used a batch size of 32 sequences with a maximum length of 512 tokens and a gradient accumulation of 2 steps, resulting in a total batch size of 64. We trained each configuration using three random seeds. The rest of hyper-parameters were left to the default values of HuggingFace's codebase (Wolf et al., 2019). The complete list of hyper-parameter values is displayed in Appendix B.

We applied this fine-tuning strategy to three different NER datasets. The first two use annotations on curated medical data (clinical cases extracted from medical literature), whereas the last one uses medical records from the ICTUSnet project.[11] More details are given below.

PharmaCoNER (Gonzalez-Agirre et al., 2019) is a track on chemical and drug mention recognition from Spanish medical texts. The authors compiled a manually classified collection of clinical case report sections derived from open access Spanish medical publications, named the Spanish Clinical Case Corpus (SPACCC). The corpus contained a total of 1,000 clinical cases and 396,988 words and was manually annotated, with a total of 7,624 entity mentions, corresponding to four different mention types.[12]

CANTEMIST (Miranda-Escalada et al., 2020) is a shared task focused on named entity recognition of tumor morphology, in Spanish. The CANTEMIST corpus[13] is a collection of 1,301 oncological case reports written in Spanish, with a total of 63,016 sentences and 1,093,501 tokens.

The **ICTUSnet** dataset consists of 1,006 hospital discharge reports of patients admitted for stroke from 18 different Spanish hospitals. It contains more than 79,000 annotations for 51 different variables. The dataset is part of the ICTUSnet project, whose main objective was the development of an information extraction system to support domain experts when identifying relevant information in discharge reports.

Finally, we remark that our main goal is a head-to-head comparison between different models to assess the best model pre-training choice. We were not aiming at maximizing results on the NER tasks and therefore we decided not to use sophisticated classification layers that might improve the performances, such as Conditional Random Field (Lafferty et al., 2001) layers on top of Bidirectional Long Short-Term Memory Recurrent Networks (Panchendrarajan and Amaresan, 2018). We argue that a simpler token classification layer better evaluates the quality of model representations than a task-specific layer. Unlike Sun and Yang (2019), where authors fine-tuned for 200 epochs (obtaining the best results using 100 epochs), we limit the fine-tuning to 20 epochs, and we do not merge the train and development sets in order to improve the results. We consider that fine-tuning for 200 epochs goes against the pre-training/fine-tuning philosophy that states that fine-tuning should be a relatively inexpensive step (Devlin et al., 2019), and also that fine-tuning for less epochs evaluates better the pre-training strategy.

5 Evaluation and Results

Each fine-tuning was executed on 4 NVIDIA V100 GPUs of 16GB VRAM. It took around 0.5, 1 and

[9] https://huggingface.co/PlanTL-GOB-ES/bsc-bio-es

[10] https://huggingface.co/PlanTL-GOB-ES/bsc-bio-ehr-es

[11] https://ictusnet-sudoe.eu/es/

[12] For a detailed description, see https://temu.bsc.es/pharmaconer/

[13] CANTEMIST corpus: https://doi.org/10.5281/zenodo.3878178

		Average of all configurations			Best on development set		
Task	Model	F1	Precision	Recall	F1	LR	Epoch
PharmaCoNER	bsc-bio-es*	$0.8907_{0.01}$	$0.8736_{0.01}$	$0.9085_{0.00}$	$0.8939_{0.01}$	5e-5	15
	bsc-bio-ehr-es*	$\mathbf{0.8913_{0.01}}$	$\mathbf{0.8758_{0.01}}$	$\mathbf{0.9073_{0.01}}$	$\mathbf{0.8954_{0.01}}$	3e-5	10
	XLM-R-Galén	$0.8754_{0.01}$	$0.8591_{0.02}$	$0.8924_{0.01}$	$0.8883_{0.00}$	5e-5	15
	BETO-Galén	$0.8537_{0.02}$	$0.8399_{0.02}$	$0.8680_{0.01}$	$0.8741_{0.01}$	5e-5	20
	mBERT-Galén	$0.8594_{0.01}$	$0.8469_{0.02}$	$0.8722_{0.01}$	$0.8760_{0.00}$	5e-5	15
	mBERT	$0.8671_{0.01}$	$0.8540_{0.02}$	$0.8809_{0.01}$	$0.8729_{0.00}$	3e-5	13
	BioBERT	$0.8545_{0.01}$	$0.8502_{0.01}$	$0.8590_{0.01}$	$0.8533_{0.01}$	2e-5	12
	roberta-base-bne	$0.8474_{0.02}$	$0.8430_{0.02}$	$0.8520_{0.02}$	$0.8680_{0.01}$	5e-5	13
CANTEMIST	bsc-bio-es*	$0.8220_{0.01}$	$0.7939_{0.02}$	$0.8522_{0.01}$	$0.8351_{0.00}$	5e-5	20
	bsc-bio-ehr-es*	$\mathbf{0.8340_{0.01}}$	$\mathbf{0.8141_{0.01}}$	$\mathbf{0.8551_{0.01}}$	$\mathbf{0.8449_{0.00}}$	5e-5	20
	XLM-R-Galén	$0.8078_{0.02}$	$0.7755_{0.02}$	$0.8431_{0.01}$	$0.8259_{0.00}$	5e-5	15
	BETO-Galén	$0.8153_{0.01}$	$0.7933_{0.02}$	$0.8387_{0.01}$	$0.8332_{0.01}$	5e-5	20
	mBERT-Galén	$0.8168_{0.01}$	$0.7919_{0.02}$	$0.8435_{0.01}$	$0.8304_{0.00}$	5e-5	20
	mBERT	$0.8116_{0.01}$	$0.7923_{0.02}$	$0.8319_{0.01}$	$0.8257_{0.00}$	5e-5	16
	BioBERT	$0.8070_{0.01}$	$0.7848_{0.02}$	$0.8306_{0.01}$	$0.8219_{0.00}$	5e-5	20
	roberta-base-bne	$0.7875_{0.03}$	$0.7733_{0.03}$	$0.8023_{0.02}$	$0.8161_{0.00}$	5e-5	15
ICTUSnet	bsc-bio-es*	$0.8727_{0.01}$	$0.8359_{0.01}$	$\mathbf{0.9131_{0.01}}$	$0.8804_{0.00}$	5e-5	19
	bsc-bio-ehr-es*	$\mathbf{0.8756_{0.00}}$	$0.8418_{0.01}$	$0.9122_{0.00}$	$0.8781_{0.00}$	2e-5	18
	XLM-R-Galén	$0.8716_{0.01}$	$0.8375_{0.01}$	$0.9087_{0.01}$	$\mathbf{0.8809_{0.00}}$	5e-5	17
	BETO-Galén	$0.8498_{0.01}$	$0.8226_{0.01}$	$0.8791_{0.01}$	$0.8551_{0.00}$	5e-5	20
	mBERT-Galén	$0.8509_{0.01}$	$0.8219_{0.01}$	$0.8820_{0.01}$	$0.8576_{0.00}$	5e-5	17
	mBERT	$0.8631_{0.01}$	$0.8301_{0.01}$	$0.8989_{0.01}$	$0.8646_{0.01}$	2e-5	20
	BioBERT	$0.8521_{0.00}$	$0.8132_{0.01}$	$0.8950_{0.01}$	$0.8503_{0.00}$	2e-5	16
	roberta-base-bne	$0.8677_{0.01}$	$\mathbf{0.8456_{0.01}}$	$0.8910_{0.01}$	$0.8769_{0.00}$	5e-5	18

Table 2: Fine-tuning results of the models for each dataset on the test set. In bold, the best results for metric and task. Subscript numbers indicate the standard deviations. Our models are marked with an asterisk.

2 hours to complete the PharmaCoNER, CANTEMIST and ICTUSnet tasks, respectively.

We then report the overall best scores on the test set, obtained by using the best model's hyper-parameters on the development set for each dataset (the standard deviation is computed using all the seeds for that configuration). Finally, we also report the models' average scores and standard deviations by computing statistics across all the seeds and the learning rates used for each dataset. The average scores are helpful to indicate which model is more robust to the variation of hyper-parameters, which are the learning rate and initial seed in our case. A higher average score and a smaller standard deviation minimizes the risk of obtaining poor results when performing an extensive hyper-parameter search is not feasible.

We compared our models with a general-domain Spanish model (`roberta-base-bne`) (Gutiérrez-Fandiño et al., 2022), a general-domain multilingual model that supports Spanish (`mBERT`), a domain-specific English model (`BioBERT`), and three domain-specific models based on continual pre-training: `mBERT-Galén` (based on mBERT), `BETO-Galén` (based on BETO, a general-domain Spanish model), and `XLM-R-Galén` (based on XLM-RoBERTa, a general-domain multilingual model supporting Spanish). The results are shown in Table 2. The last two columns report the learning rate and epoch in which the best configuration on the development set was achieved

Our models obtained significantly better performances than the general-domain models, namely `mBERT` and `roberta-base-bne`. Compared to the domain-specific Galén models, our average models' scores surpassed them on the clinical NER tasks. However, when looking at the best on development score on the ICTUSnet dataset, the `XLM-R-Galén` model outperformed our models. We also highlight that our models exhibit smaller standard deviations. This makes them more robust and a good option if not enough computational

resources are available to experiment with the different hyper-parameter configurations.

6 Conclusions and Future Work

This work presents the first large-scale biomedical Spanish language models trained from scratch, using a large biomedical corpora for a total of 1.1B tokens and an EHR corpus of 95M tokens. We fine-tuned the models on three clinical NER tasks and compared them with both general-domain and other available Spanish clinical models. The results show the superiority of our models across the NER tasks, making them competitive candidates for clinical NLP applications. Our findings demonstrate the benefits of pre-training from scratch, as seen in Gu et al. (2021). Regarding continual pre-training, the benefits are not clear, especially when continual pre-training is performed with small data, as in the case of the `mBERT-Galén`, `XLM-R-Galén`, and `BETO-Galén` (note that `mBERT` outperforms `mBERT-Galén` in two out of three tasks). Our work raises exciting research questions about which pre-training approach is optimal to tackle challenging clinical NLP tasks. We will devote future efforts to address the previous question in detail by providing new models based on continual pre-training and extending our evaluation setting to a diverse range of tasks.

7 Data Availability

Our work encourages the development of Clinical and Biomedical NLP applications for Spanish. Therefore, we released our pre-trained models and the best on dev set fine-tuned models under the Apache License 2.0 in the HuggingFace models hub under the following links:

Pre-trained models

- bsc-bio-es
- bsc-bio-ehr-es

Fine-tuned models

- bsc-bio-ehr-es-pharmaconer
- bsc-bio-ehr-es-cantemist

Moreover, to guarantee reproducibility, we share the script used to fine-tuned our pre-trained model in the official GitHub repository: `https://github.com/PlanTL-GOB-ES/lm-biomedical-clinical-es`.

Acknowledgements

This work was funded by the Spanish State Secretariat for Digitalization and Artificial Intelligence (SEDIA) within the framework of the Plan-TL.[14]

[14] `https://plantl.mineco.gob.es/Paginas/index.aspx`

References

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.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063.

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.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Ona de Gibert Bonet, Asier Gutiérrez-Fandiño, Aitor Gonzalez-Agirre, Martin Krallinger, and Marta Villegas. 2021. Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

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.

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.

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 Trans. Comput. Healthcare*, 3(1).

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Spanish legalese language model and corpora.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

John D. Lafferty, Andrew McCallum, and Fernando 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.

Jieh-Sheng Lee and Jieh Hsiang. 2019. Patentbert: Patent classification with fine-tuning a pre-trained BERT model. *CoRR*, abs/1906.02124.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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. *CoRR*, abs/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.

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*.

Aurélie Névéol, H. Dalianis, G. Savova, and Pierre Zweigenbaum. 2014. Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of Biomedical Semantics*, 9.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. Bidirectional LSTM-CRF for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

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, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking. *CoRR*, cs.CL/0009008.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Cong Sun and Zhihao Yang. 2019. Transfer learning in biomedical named entity recognition: An evaluation of BERT in the PharmaCoNER task. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 100–104, Hong Kong, China. Association for Computational Linguistics.

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

A Pre-training Hyper-parameters

The hyper-parameters used for pre-training our models are shown in Table 3.

Hyper-parameter	Value
Number of Layers	12
Hidden size	768
FNN inner hidden size	3072
Attention Heads	12
Attention Head size	64
Dropout	0.1
Attention Dropout	0.1
Warmup Steps	10k
Peak Learning Rate	5e-4
Batch Size	2,048
Weight Decay	0.01
Max Steps	125k
Learning Rate Decay	Linear
Adam ϵ	1e-6
Adam β_1	0.9
Adam β_2	0.98
Gradient Clipping	0.0

Table 3: Hyper-parameters used for pre-training.

B Fine-tuning Hyper-parameters

The hyper-parameters used for fine-tuning the models on various tasks are shown in Table 4.

Hyper-parameter	Value
Learning Rates	{0.8, 1, 2, 3, 5}e-5
Learning Rate Decay	Linear
Warmup Steps	0
Batch Size	64
Weight Decay	0.0
Max. Training Epochs	20
Adam ϵ	1e-8
Adam β_1	0.9
Adam β_2	0.999
Gradient Clipping	1.0

Table 4: Hyper-parameters used for fine-tuning.