

Coming a Long Way with Pre-Trained Transformers and String Matching Techniques: Clinical Procedure Mention Recognition and Normalization

Notebook for the BioASQ Lab at CLEF 2023

Mariia Chizhikova^{1,*†}, Jaime Collado-Montañez^{1,†}, Manuel Carlos Díaz-Galiano¹, L. Alfonso Ureña-López¹ and María Teresa Martín-Valdivia¹

¹Department of Computer Science, University of Jaén, Campus Las Lagunillas, s/n, Jaén, 23071, Spain

Abstract

This paper covers the participation of the SINAI team in the MedProcNER shared task and the BioASQ workshop held on CLEF 2023. The main objective of this challenge is to create systems able to accurately detect and normalize clinical procedure mentions in electronic health Reports. For the named entity recognition (NER) sub-task we compare different ways of processing long sequences: sentence level token classification based on fine-tuning of a RoBERTa model pre-trained on biomedical and clinical data and employing different types of recurrent architectures that rely on non-trainable contextual word embeddings extracted from the same pre-trained language model. In the normalization sub-task, we perform a sequential process that combines literal string matching and embedding similarity search to link each entity found in the previous sub-task with a concept from the SNOMED-CT ontology. Our best-performing system achieves 0.7568 micro-averaged F1 score on the NER sub-task and 0.5267 on the NORM sub-task.

Keywords

Clinical NLP, Named Entity Recognition, Electronic Health Records

1. Introduction

Automated coding and classification technologies comprise diverse computer-based methodologies aimed at converting unstructured narrative text found in clinical records into structured text. These methodologies involve the assignment of codes derived from standard terminologies, all without requiring human intervention [1]. Structuring clinical information contained in free-text clinical narratives enables a large variety of applications including assistance of healthcare professionals with retrospective studies and clinical decision-making support [2]. Structured clinical information at a large scale can also be leveraged by medical and pharmacologic inquiries

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.


†These authors contributed equally.

✉ mchizhik@ujaen.es (M. Chizhikova); jcollado@ujaen.es (J. Collado-Montañez); mcdiaz@ujaen.es (M. C. Díaz-Galiano); laurena@ujaen.es (L. A. Ureña-López); maite@ujaen.es (M. T. Martín-Valdivia)

🆔 0000-0002-0302-912X (M. Chizhikova); 0000-0002-9672-6740 (J. Collado-Montañez); 0000-0001-9298-1376 (M. C. Díaz-Galiano); 0000-0001-7540-4059 (L. A. Ureña-López); 0000-0002-2874-0401 (M. T. Martín-Valdivia)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

with the aim of efficiently bringing clinical evidence to medical research.

Identification of clinically relevant terms in patients' electronic health records (EHR) and their mapping to codes from a relevant controlled vocabulary is a time-consuming task that entails errors attributed to the human factor. For this reason, the automation of clinical coding had attracted the attention of the NLP community. Since the field of clinical information retrieval was formed, a large variety of approaches has been proposed to tackle this task ranging from early rule-based methods [3] to sophisticated applications based on deep learning [4].

Regarding the relevant entities being detected, extraction of disease and drug mentions might be the most investigated field within this specific branch of clinical NLP: many shared tasks brought the community effort to it by making available high-quality datasets with expert-generated annotations [5, 6, 7].

Clinical procedures encompass a range of activities undertaken by healthcare professionals to diagnose, treat, or manage a patient's medical condition. These activities, such as physical examinations, laboratory tests, imaging studies, surgical procedures, medication administration, and other medical interventions, are vital in the realm of patient care. Their significance lies in aiding healthcare providers in the diagnosis and treatment of medical conditions, as well as monitoring disease progression and averting complications. Furthermore, clinical procedures contribute significantly to the advancement of medical knowledge and the enhancement of healthcare outcomes through research endeavors.

By providing a dataset of reports extensively annotated by experts, MedProcNER task [8] at the BioASQ workshop [9] fosters collective collaboration within the clinical NLP community in the development of systems capable of correctly identifying and mapping to SNOMED-CT¹ codes of procedure mentions in reports written in Spanish.

This paper presents the contribution of the SINAI team to the MedProcNER shared task. We participated in two sub-tasks: the first involving the detection of procedure mentions in clinical reports and the second consisting in mapping the terms detected with the named entity recognition (NER) system to codes from the SNOMED-CT ontology. We refer to these sub-tasks as NER and NORM respectively.

2. Data

The MedProcNER corpus comprises 1,000 clinical case reports in Spanish, which have been annotated with mentions of clinical procedures and normalized to SNOMED-CT codes [8]. These texts are sourced from the SPACCC corpus [10] and are identical to the ones utilized in the DisTEMIST [11] shared task at BioASQ 2022, thus making the annotations mutually beneficial for medical entity recognition. In this section we provide a brief exploratory-descriptive analysis of the corpus and describe the pre-processing procedures applied to prepare this data for being fed to the systems described in Section 3.

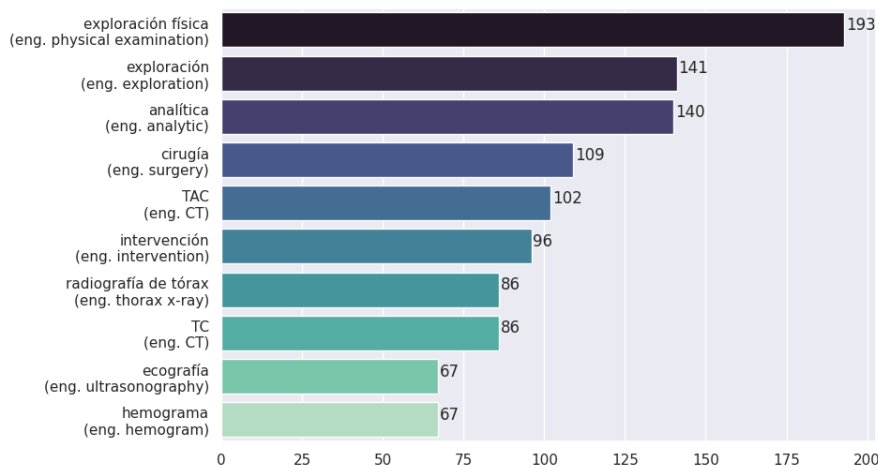
¹https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html

2.1. Exploratory descriptive analysis

In order to make possible a competitive evaluation of the system presented by teams on the task, the dataset was split by organizers into training and test subsets, the former containing 750 reports and being released with all the annotations and the latter totaling 250 reports and being released in plain text format.

Each report in the training set was annotated with at least one procedure mention, reaching 48 as the maximum number of annotations with an average of 14.77 procedure mentions per text, and being the standard deviation equal to 8.3807. Figure 1 displays the 10 most frequently observed entities across the training data.

Figure 1: 10 most frequently observed entity spans across the training data subset. *English translation made only to ease the reading*



As for the length of the text in the MedProcNER corpus, Table 1 summarizes the related statistics obtained by splitting the text with the RoBERTa Byte-Pair Encoding tokenizer [12], the same we employ in the proposed systems. It can be noted that there are no relevant differences between the two provided subsets in terms of this characteristic.

Table 1
Corpus length statistics

Subset	Maximum length (tokens)	Minimum length (tokens)	Avg (STD)
Train	1486	98	458.05 (218.16)
Test	1439	115	458.14 (233.24)

2.2. Pre-processing

In order to make the data format coherent with the approaches we selected to tackle both sub-tasks we developed different pre-processing procedures.

To be able to evaluate our systems during the development process, we divided the training set provided by the organizers into two subsets: one that contained 80% of the data (601 reports) was selected for training, and 20% (149 reports) was reserved for performing the in-house evaluation.

As will be covered in detail in Section 3, the application of transformer models to the token classification task poses the challenge related to the maximum input length restrictions. This led us to employ three different pre-processing procedures.

For the first approximation, we split the reports into sentences using the *SentenceRecognizer* from the SpaCy processing pipeline². SpaCy’s *SentenceRecognizer* relies on `es_core_news_sm` pre-trained language model³ which predicts whether each token of every text starts a sentence or not. For the second approximation, we truncated all long texts to 512 tokens. Finally, the third approximation consisted in no applying any pre-processing at all.

As for the labeling scheme, we performed experiments with two different types: BIO and BIOES. In the BIO scheme the ‘B-’ prefix of the label indicates that the token corresponds to the beginning of the named entity, ‘I-’ prefix marks the labels of tokens inside of the entity while ‘O’ label is reserved for tokens that do not form part of any named entity. BIOES scheme is a little bit more sophisticated because it also distinguishes the end (‘E-’ prefix) of an annotation and single-token entities (‘S-’ prefix).

3. System description

In this section we describe the systems presented for the official evaluation on the two sub-tasks.

3.1. Sub-task 1: NER

The advent of large language models like BioBERT [13] or BioM-ELECTRA [14] has proven the benefits of domain-specific pre-training that enables such models to capture the contextual representation of the corpora thus improving the performance of these systems on downstream tasks such as text classification or NER. For this reason, we opted for basing our system on `bsc-bio-ehr-es`, a RoBERTa architecture model pre-trained on a combination of biomedical and clinical corpora [15]. This system achieved state-of-the-art (SOTA) performance on Spanish clinical NLP benchmarks like CANTEMIST (tumor morphology mentions extraction) [16] and PharmaCoNER (pharmacological substance mention extraction) [17].

One of the main challenges in using contextual representations from such transformer architectures as BERT or RoBERTa is the limit of the maximum length of the input text, which is set to 512 tokens.

As we stated in Section 2, despite the fact of the average length of clinical reports in the corpus being below the maximum length threshold of the `bsc-bio-ehr-es` model, the maximum length exceeds 1400 tokens in both subsets. To be more precise, 248 texts from the train set (33.07% of the total) and 87 texts from the test set (34.8% of the total) are more than 512 tokens long. Moreover, the tendency of longer reports to contain more annotation makes to seem

²<https://spacy.io/api/sentencerecognizer>

³<https://spacy.io/models/es>

unfeasible the widely adopted approach of text truncation. More specifically, truncation of the texts from training and validation data subsets entails a loss of 49.5% of all the annotations, lowering the mean number of entities per document from 14.77 to 7.14.

Unlike the SOTA approach to transformer model fine-tuning by adding a classification head typically consisting of a linear layer combined with some regularization techniques like dropout [18], using recurrent classifiers like Long Short-Term Memory (LSTM) layers or Gated Recurrent Unit (GRU) layers doesn't require the inputs to have the same dimensions and a priori do not have maximum input length restrictions [19, 20]. Recently, approaches that combine the benefits of transformers and recurrent models in order to improve NER quality were proposed [21]. Nonetheless, the recurrent nature of such classifiers makes them particularly vulnerable to the vanishing gradient problem, especially when dealing with very long sequences.

With the objective of evaluating the viability of a combination of transformer embeddings with recurrent classifiers of different architectures, our team submitted a total of 5 runs for the NER sub-task.

The first system (run-1) corresponded to a sentence level token classification approach that used the BIO scheme for entity labeling and relied upon fine-tuning the `bsc-bio-ehr-es` pre-trained model by adding a 0.1 dropout and a linear layer on top of the original RoBERTa architecture. In order to maximize the performance of the fine-tuned model we selected the values for learning rate, training batch size, weight decay, Adam optimizer epsilon and the number of warm up steps after 5 trials of hyperparameter optimization that relied on the Optuna framework [22]. The search space and the selected value for each of the parameters are summarized in Table 2. To determine the number of training epochs we implemented an early stopping strategy that stopped training after 3 epochs without improvement of the F1-score during the evaluation of the development set.

Table 2
Hyperparameter optimization search space and results

Parameter	Search space	Selected Value
Learning rate	Float value between $3e - 5$ and $5e - 5$	$3.5e - 5$
Training batch size	Either 8 or 16	8
Weight decay	Float value between $1e - 12$ and $1e - 1$	$1.8e - 3$
<i>Adame</i>	Float value between $1e - 10$ and $1e - 6$	$1.1e - 9$
Warm up steps	integer value between 0 and 1000	496
Training epochs	-	11

The second system (run-2) is a classifier with one layer of Bidirectional LSTM (Bi-LSTM) followed by a Linear layer and a Conditional Random Field (CRF) layer. This classifier takes as input non-trainable representations of texts truncated to the maximum length of 512 tokens and labeled following the BIOES scheme. Those representations were obtained as a mean of the output from the last four layers of the same `bsc-bio-ehr-es` pre-trained model, but without subjecting it to any fine-tuning. We utilized a sliding window technique to obtain a contextual representation for each token, adjusting the size of the window to 128. In other words, each token embedding corresponds to its representation in a context of 64 preceding and 64 following tokens. As in the previously described approach, we evaluated the system on the

Table 3

Summary of the main distinctive features of the systems presented

	run-1	run-2	run-3	run-4
Maximum length	sentence level	512	full text	full text
Labeling scheme	BIO	BIOES	BIOES	BIO
Trainable embeddings	True	False	False	False
Contextual embeddings	False	True	True	True
Classifier	Linear	LSTM+CRF	GRU+CRF	LSTM+CRF

development set after each epoch and interrupted training after 3 epochs without improvements of the reference metric, which occurred at the end of 109th.

The third system (run-3) generated text representations following the same approach as described for run-2, but this time we did not truncate the text of the reports. The labeling scheme used in this case was BIOES. As the core layer of the trained classifier, we selected a Bi-GRU layer, as a popular and more lightweight alternative for the Bi-LSTM. This layer was followed by a Linear layer and a CRF. The early stopping callback interrupted this system’s training process after 106 epochs of training.

The fourth system (run-4) is a modification of the previous one, with a Bi-LSTM layer instead of a Bi-GRU. This system started to overfit more quickly, as the training was stopped after completing epoch 87.

Finally, the fifth presented system (run-5) was identical to the run-4 but used BIO labeling scheme rather than BIOES.

Table 3 provides a summary of the distinctive characteristics of all four presented systems.

All systems were trained on a single NVIDIA A-100 40GB GPU by making use of Huggingface transformers Python library [23] for the first run and Flair Python toolkit [24] for the other ones.

3.2. Sub-task 2: NORM

The normalization sub-task requires linking every detected mention during the previous task to a SNOMED CT concept. To this end, we developed three different linkers which are executed sequentially.

As an initial step, we make use of the golden labels from the training set provided by the organizers. We perform a look-up matching process where any entity found exactly in this set is assigned its golden code. After this, all the entities that were not found in the previous step follow a similar matching, assigning a SNOMED-CT code to any entity exactly present in the given gazetteer. Finally, we precalculate an embedding vector for every term in the gazetteer using SapBERT-XLMR [25], which is a transformer model trained with UMLS 2020AB. We also embed all the remaining unannotated entities from the NER subtask. Then, we perform a similarity search between both embedding arrays by using the Faiss library [26]. As a result of this last process, each entity gets assigned the code of its most similar term in the ontology.

Table 4

Official results obtained by the SINAI team in Clinical Procedure Recognition and Normalization subtasks along with the best-performing system in the competition.

Subtask	System	MiP	MiR	MiF1
NER	run 1	0.7631	0.7505	0.7568
	run 2	0.7786	0.7043	0.7396
	run 3	0.7396	0.7110	0.7250
	run 4	0.7538	0.7353	0.7444
	run 5	0.7705	0.7049	0.7362
NORM	run 1	0.5310	0.5224	0.5267
	run 2	0.5455	0.4936	0.5183
	run 3	0.5079	0.4884	0.4980
	run 4	0.5173	0.5047	0.5109
	run 5	0.5352	0.4898	0.5115

4. Results

In this section, we present the results obtained by the systems we developed as part of our participation in MedProcNER sub-tasks 1 and 2 at BioASQ 11. The evaluation metrics for both tasks, which are computed by comparing the generated predictions to the expert’s manual annotations, are micro-averaged precision (MiP), micro-averaged recall (MiR) and micro-averaged F1-score (MiF), being the latter the reference metric for the leader board.

Table 4 displays the evaluation metrics scored by all the presented systems during the official evaluation of the test set.

On the NER sub-task, the best performing system in terms of MiF resulted to be run-1, the fine-tuned RoBERTa system that approached the task of procedure mention recognition as a sentence level token classification task. Nevertheless, the incorporation of a recurrent classifier based on the LSTM layer was proven to be beneficial for the system’s precision, making them suitable for environments where a low false positive rate is crucial.

As sub-task 2 relies on the results obtained in the NER task, in Table 4 we show the NORM scores of the described system applied to the predictions made by every NER system developed. The best-performing system is the one applied to the entities recognized in run 1 from the first sub-task, which achieved a MiP of 0.5310, a MiR of 0.5224, and a MiF1 of 0.5267.

In Table 5 we describe the impact of every step in our system as the number of entities matched during the first and second steps of our sequential approach. Our best-performing system in the first sub-task found a total of 3559 entities in the test set, which is the highest number across all the runs. Run 4 extracted 3531 entities, and 50% of them were successfully matched with either the train set or the ontology. The second system found a relatively low number of entities with respect to the other runs (3275), making the exact matching technique more relevant in the overall score for this experiment as a higher percentage of the codes assigned in this manner has been achieved (52.18%). This is also reflected in the fact that this run scored the highest MiP in the test set (0.5455).

Table 5

Entities exactly matched in steps 1 and 2 during sub-task 2.

System	Total entities	Step 1 matches	Step 2 matches	% Exact matches
run 1	3559	1632	147	49.99%
run 2	3275	1546	163	52.18%
run 3	3479	1515	186	48.89%
run 4	3531	1580	185	49.99%
run 5	3311	1510	175	50.89%

5. Conclusions and future work

In this paper we present the system developed by the SINAI team at the MedProcNER shared task to tackle the sub-tracks of clinical procedure mention recognition and normalization.

For the NER subtask, we compare four different systems with the aim of evaluating different approaches to processing long sequences: we experiment with performing sentence level token classification, text truncation and full text processing. To test the utility of transformer embeddings generated in a contextual manner we employed a set of recurrent classifiers that operate with different labeling schemes and various maximum text lengths.

All five presented systems showed promising and close to each other results. In terms of MiF, a sentence level NER system based on a RoBERTa model pre-trained on biomedical and clinical corpora and fine-tuned by adding a dropout and a linear layer on top of the original architecture resulted to be the best performing one with a 0.7568. The fact that the lowest false positive rate was shown by an LSTM classifier that operated with a contextual representation of reports truncated to the maximum length of 512 tokens suggests the need of performing sentence level NER employing the same recurrent architecture to accurately measure the benefits of this approach. We also leave for future work the ablation study of such parameters as the number of layers used to generate the contextual representation of text that is fed to the recurrent classifier and the impact of adding the CRF layer.

Regarding the NORM subtask, we performed a sequential matching where we first look for the entities in the training set, then we match the remaining unannotated entities against the SNOMED-CT gazetteer provided by the organizers, and finally, we compute a similarity search between the embedded ontology terms and the entities that were not found in the previous steps. As this task relies on the results achieved in the previous task, the best-performing system has been run 1 too, scoring a micro-averaged F1-score of 0.5267. As for future work, we aim to try different techniques for the entities that could not be found in the exact matching process such as data augmentation to expand the pool of potential candidates in the gazetteer or the use of fuzzy techniques such as Levenshtein distance to fix possible grammar mistakes in the input files.

Acknowledgments

This work has been partially supported by WeLee project (1380939, FEDER Andalucía 2014-2020)

funded by the Andalusian Regional Government, and projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government, and project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government.

References

- [1] M. H. Stanfill, M. Williams, S. H. Fenton, R. A. Jenders, W. R. Hersh, A systematic literature review of automated clinical coding and classification systems, *Journal of the American Medical Informatics Association* 17 (2010) 646–651.
- [2] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical natural language processing in languages other than english: opportunities and challenges, *Journal of biomedical semantics* 9 (2018) 1–13.
- [3] A. M. Scott, Automatic coding of a diagnosis, in: G. McLachlan, R. Shegog (Eds.), *Computers in the Service of Medicine*, volume 2, Oxford University Press London, 1968, p. 89.
- [4] P. López-Úbeda, M. C. Díaz-Galiano, L. A. Ureña-López, M. T. Martín-Valdivia, Combining word embeddings to extract chemical and drug entities in biomedical literature, *BMC bioinformatics* 22 (2021) 1–18.
- [5] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, A. Valencia, Overview of the chemical compound and drug name recognition (chemdner) task, in: *BioCreative challenge evaluation workshop*, volume 2, Citeseer, 2013, p. 2.
- [6] A. Gonzalez-Agirre, M. Marimon, A. Intxaurreondo, O. Rabal, M. Villegas, M. Krallinger, PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track, in: *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1–10. URL: <https://aclanthology.org/D19-5701>. doi:10.18653/v1/D19-5701.
- [7] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020., *CLEF (Working Notes) 2020* (2020).
- [8] S. Lima-López, E. Farré-Maduell, L. Gascó, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023, in: *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*, 2023.
- [9] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, L. Gasco, M. Krallinger, G. Paliouras, Overview of BioASQ 2023: The eleventh BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*, ????
- [10] A. Intxaurreondo, M. Marimon, A. Gonzalez-Agirre, J. A. Lopez-Martin, H. Rodriguez, J. Santamaria, M. Villegas, M. Krallinger, Finding mentions of abbreviations and their definitions in spanish clinical cases: The barr2 shared task evaluation results., *IberEval@SEPLN 2150* (2018) 280–289.
- [11] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis,

- A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2022.
- [12] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1715–1725. URL: <https://aclanthology.org/P16-1162>. doi:10.18653/v1/P16-1162.
- [13] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* 36 (2020) 1234–1240.
- [14] S. Alrowili, V. Shanker, BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 221–227. URL: <https://aclanthology.org/2021.bionlp-1.24>. doi:10.18653/v1/2021.bionlp-1.24.
- [15] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in Spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: <https://aclanthology.org/2022.bionlp-1.19>. doi:10.18653/v1/2022.bionlp-1.19.
- [16] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results., *IberLEF@ SEPLN* (2020) 303–323.
- [17] A. Gonzalez-Agirre, M. Marimon, A. Intxaurrenondo, O. Rabal, M. Villegas, M. Krallinger, Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track, in: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 1–10. URL: <https://aclanthology.org/D19-5701>. doi:10.18653/v1/D19-5701.
- [18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423>. doi:10.18653/v1/N19-1423.
- [19] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural computation* 9 (1997) 1735–1780.
- [20] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).
- [21] J. Li, T. Wang, W. Zhang, An improved chinese named entity recognition method with tb-lstm-crf, in: 2020 2nd Symposium on Signal Processing Systems, 2020, pp. 96–100.
- [22] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperpa-

- parameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.
- [23] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.
- [24] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, 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), 2019, pp. 54–59.
- [25] F. Liu, I. Vulić, A. Korhonen, N. Collier, Learning domain-specialised representations for cross-lingual biomedical entity linking, in: Proceedings of ACL-IJCNLP 2021, 2021, pp. 565–574.
- [26] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Transactions on Big Data 7 (2019) 535–547.