# Biomedical Spanish Language Models for entity recognition and linking at BioASQ DisTEMIST

Vincenzo Moscato[1,2], Marco Postiglione[1] and Giancarlo Sperlí[1,2]

[1]University of Naples Federico II, Department of Electrical Engineering and Information Technology (DIETI), Via Claudio, 21 - 80125 - Naples, Italy

[2]Consorzio Interuniversitario Nazionale per l'Informatica (CINI) - ITEM National Lab, Complesso Universitario Monte S.Angelo, Naples, Italy

**Abstract**

Named Entity Recognition and Entity Linking systems usually require a rich and annotated dataset to be trained and produce high-quality results, but the annotation process is time consuming and expensive, especially when it needs the effort of domain experts, such as in the medical field. However, recent developments in Natural Language Processing (NLP) allow us to easily use transformer language models which have been pre-trained on a huge quantity of data (often coming from specialized domains), and thus obtain high performance without excessive efforts. In this work, we outline our approach to NER and EL tasks on Spanish clinical notes for the *DisTEMIST* track at the *BioASQ 2022* challenge. Our results demonstrate that the proposed methodology based on biomedical pre-trained language models turned out the best for the NER task with a $\sim 3\%$ higher $F1$ w.r.t. the second-best solution.

**Keywords**

Biomedical Named Entity Recognition, Entity Linking, Transformers, EHRs

## 1. Introduction

Biomedical Named Entity Recognition (NER) and Entity Linking (EL) are often the first and essential steps in many text understanding applications [1], such as the construction and analysis of structured knowledge bases (e.g. knowledge graphs) or conversational agents including medical chatbots and research assistants.

NER aims at recognizing mentions of pre-defined entity types within unstructured text data, while EL links them to *concepts* in a (usually) external knowledge base (e.g. UMLS [2], SNOMED CT[1]). These two tasks are the subject of the *DisTEMIST (DISease TExt Mining Shared Task)* track at the *BioASQ 2022* challenge [3], which invites researchers, biomedical industry professionals, NLP, and ontology experts to develop systems capable of indexing the content about **diseases** within Spanish clinical notes. In this work, we describe the approach of our team (*PICUSLab*) which allowed us to win the NER track with a $\sim 3\%$ higher $F1$ measure w.r.t. the second-best solution.

*CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy*

vincenzo.moscato@unina.it (V. Moscato); marco.postiglione@unina.it (M. Postiglione); giancarlo.sperli@unina.it (G. Sperlí)

http://wpage.unina.it/vmoscato/ (V. Moscato); http://wpage.unina.it/giancarlo.sperli/ (G. Sperlí)

0000-0002-0754-7696 (V. Moscato); 0000-0001-6092-940X (M. Postiglione); 0000-0003-4033-3777 (G. Sperlí)



[1]https://www.nlm.nih.gov/healthit/snomedct/index.html

The development of a biomedical text understanding system with high precision and recall is a challenging task due to the fact that biomedical datasets are characterized by a large number of *synonyms*, *alternate spellings* of entities, which are often referred to with *non-standard abbreviations*, and *polysemous words*, i.e. words that can have different meanings based on their context. For example, *VHL* may refer to the *Von Hippel-Lindau* disease or to the gene name which causes the disease, based on the context in which it appears.

Initially, NER and EL systems were mainly *dictionary*- and *rule*-based [4], but they failed dealing with unseen and polysemous words [5]. The availability of annotated datasets allowed systems to evolve by means of deep learning techniques, such as Bidirectional Long-Short Term Memory (BiLSTM) networks [6, 7] and Transformer architectures [8, 9]. However, due to the above-mentioned problems related to biomedical corpora, directly applying state-of-the-art NLP methodologies to biomedical text mining has limitations, since language models are trained and tested mainly on datasets containing general domain texts (e.g. Wikipedia). Hence, recent models proposed in biomedical text mining rely on adapted versions of pre-trained language models [10, 11], even in low- and mid-resource languages [12, 13].

In this work, we describe our approach to the DisTEMIST track, which is based on embeddings computed with a Spanish biomedical RoBERTa backbone network [13]. We use a simple classification head (a linear layer with a softmax activation function) to produce an high-quality NER system, and then apply a similarity-based EL process to entities retrieved in the NER step. Our experimental results show the appropriateness of the methodology.

## 2. Tasks formulation

We start with a corpus of annotated sentences $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{X} \times \mathcal{Y}\}$, where:

- $i \in \{1, ..., N\}$, $N$ is the length of the dataset
- $\mathcal{X}$ is the set of sentences
- $\mathbf{x}_i$ is a sentence, which is defined as a sequence of tokens $x_j \in \mathbf{x}_i$, $j \in \{1, \ldots, H_i\}$, where $H_i$ is the sequence length
- $\mathcal{Y}$ is the set of labels. In our work, we will refer to the IOB2 annotation scheme [14], thus $\mathcal{Y} = \{B, I, O\}$, where $B$ indicates the *beginning*, $I$ the *inside* and $O$ the *outside* of an entity mention
- $\mathbf{y}_i$ maps each token $x_j \in \mathbf{x}_i$ to its corresponding label $y_j$.

Based on this corpus, the objective of a NER model is to assign the correct label in $\mathcal{Y}$ to each token in an input sentence.

Given the set of entity mentions $M$ resulting from the NER step, and a knowledge base containing a set of entities $E$, EL aims to map each entity mention $m \in M$ to the most appropriate concept $e \in E$.

## 3. Materials

The DisTEMIST **corpus** was manually annotated by clinical experts following guidelines containing rules for annotating diseases in Spanish clinical cases. Guidelines were created de

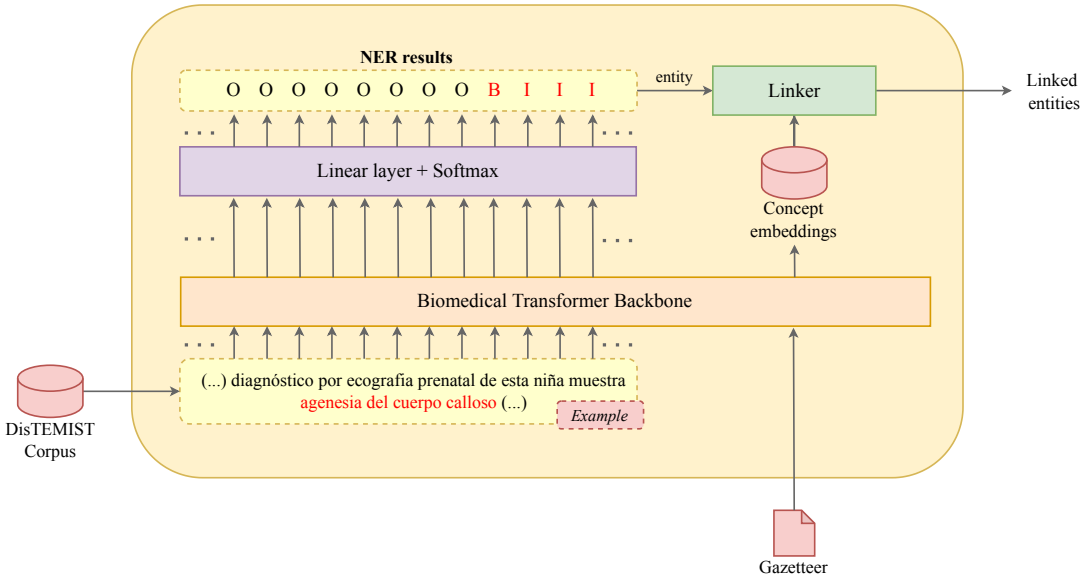


**Figure 1:** Overview of our NER + EL solution for the DisTEMIST track. A biomedical Spanish pre-trained Transformer backbone network is used to compute: (1) *token embeddings* to be classified by a classification head (linear layer + softmax); (2) *concept embeddings* for each concept within the *gazetteer* which will be used by the *Linker* to associate the nearest concept to an entity mention based on similarity measures.

novo by clinical experts defined after several cycles of quality control and annotation consistency analysis before annotating the entire dataset.

The training set for NER and EL consists of 750 and 584 annotated clinical cases, respectively. However, every entity mention in the original DisTEMIST corpus has been linked to a Snomed-CT, also when the exact concept is not present within the ontology (e.g. "Chorioretinal lacunae" is normalized to "Chorioretinal disorder").

The DisTEMIST **gazetteer** contains main terms and synonyms from the relevant branches of Snomed-CT for the grounding of disease mentions.

## 4. Methodology

Figure 1 shows an overview of the methodological flow of our solution for the DisTEMIST track. A Transformer backbone network pre-trained with Spanish biomedical corpora has been used in both NER and EL tasks. In the former case, it has been used to compute *token embeddings* for a classification head with a linear layer and a softmax activation function; in the latter, it computes *concept embeddings* for each concept within the *gazetteer*, which will be then used to link an entity mention to the nearest concept based on a measure of similarity. In this section, each module of our methodology will be extensively described.

### 4.1. Biomedical Transformer Backbone network

The Biomedical Transformer Backbone network used in this work has been pre-trained and made publicly available by Carrino et al. [13]. It uses a RoBERTa [15] base model with 12 self-attention layers with masked language modeling as the pre-training objective. The dataset used to pre-train the network consists in two corpora with different sizes and domains:

- *Clinical corpus*: it contains 91M tokens from more than 278K clinical documents (e.g. discharge reports, clinical course notes).
- *Biomedical corpus*: it contains data from a variety of sources, such as medical crawlers, PubMed[2] and Scielo[3] publications and patents. The entire corpus counts a total of 968M words.

### 4.2. Named Entity Recognizer

The Transformer-based backbone network is used to extract an embedded representation of each token $x_j$ in an input sample $\mathbf{x}$, $\mathbf{z} = f_{\theta_{LM}}(x_j)$, $\theta_{LM}$ being the set of language model parameters. Thereafter, a linear layer (a.k.a. *classification head*) with parameters $\theta_L = \{\mathbf{W}, \mathbf{b}\}$ project the Transformer-based representation $\mathbf{z}$ into the label space, $f_{\theta_L}(\mathbf{z}) = Softmax(\mathbf{W}\mathbf{z} + \mathbf{b})$. The model parameters are then optimized by minimizing cross-entropy:

$$\mathcal{L}_{CE} = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \sum_{i=1}^{H} KL\Big(y_i \Big| q(y_i|x_i)\Big), \tag{1}$$

where $KL(p|q)$ is the Kullback-Leibler divergence between the two distributions $p$ and $q$, and $q$ is the prediction probability vector for each token:

$$q(y|x) = Softmax(\mathbf{W} \cdot f_{\theta_{PLM}}(x) + \mathbf{b}) \tag{2}$$

### 4.3. Entity Linker

Inspired by Kraljevic et al. [16], our EL approach relies on a *Concept Database ($CDB$)* component, i.e. a table representing a concept dictionary. To this end, e used the gazetteer provided by DisTEMIST track organizers. Even though not every concept within the gazetteer appears in our training set, we decided to keep all the concepts due to the unpredictability of concepts in the test set. Our linking approach is based on *context similarity*: we learn an embedded representation for each concept and for new documents, when an entity mention is detected by the NER model, its context is compared to the embedded representations of all the concepts in the $CDB$ to choose the most appropriate one.

[2] https://pubmed.ncbi.nlm.nih.gov
[3] https://scielo.org

#### Concept Embeddings

We learn concept embeddings in a supervised fashion. For each concept $c \in CDB$, we perform the steps described as follows to compute its concept embedding $V^c_{concept}$:

1. *Initialization*: given the concept name $c_{name}$ and its description $c_{description}$ provided with the gazetteer, we initialize $V^c_{concept}$ with the embedding of the concatenation of the two strings $[c_{name}, c_{description}]$ computed with the Biomedical Transformer backbone network.
2. *Context embeddings*: for each entity in the training set annotated with the concept $c$, we compute its context embedding $V_{context}$ with the Biomedical Transformer backbone network.
3. *Update*: for each entity in the training set annotated with the concept $c$, the concept embedding $V^c_{concept}$ is updated with the context embedding $V_{context}$. Specifically, the update criterion is described by the following equation:

$$V^c_{concept} = V^c_{concept} + lr \cdot (1 - sim) \cdot V_{context}, \tag{3}$$

where:

- $lr$ is the *learning rate*, computed as $lr = \frac{1}{\mathcal{N}_c}$, $\mathcal{N}_c$ being the number of times the concept appears during training.
- $sim$ is the cosine similarity between $V^c_{concept}$ and $V_{context}$,

$$sim = max\Big(0, \frac{V^c_{concept}}{\|V^c_{concept}\|} \cdot \frac{V_{context}}{\|V_{context}\|}\Big) \tag{4}$$

#### Linking

Given the entity mention recognized by the NER model, we compute its context embedding $V_{context}$ by means of the Biomedical Transformer backbone network. Then, we compute its cosine similarity $sim$ with all the concept embeddings $V_{concept}$. We eventually link the entity with the most similar concept.

## 5. Experiments

The performance of our proposed approaches for NER and EL has been evaluated by participating to the *DISease TExt Mining Shared Task (DisTEMIST)* track within the BioASQ 2022 challenge. In this section we show the performance results of our methodology on the final test set and some preliminary experiments on the training corpus provided by the challenge organizers.

### 5.1. Experimental setup

#### Evaluation Metrics

Evaluation is done by comparing the automatically generated results to the results generated by manual annotation of experts. The primary evaluation metric for both the NER and EL sub-tracks will consist of micro-averaged precision (*MiP*), recall (*MiR*) and F1-scores (*MiF1*).

### Configuration

Both the NER and EL models were implemented using the HuggingFace Transformers library (v4.4.0) [17]. The biomedical Spanish Transformer backbone network has been downloaded from the HuggingFace model repository (`PlanTL-GOB-ES/roberta-base-biomedical-clinical-es`). To deal with the limited length of input samples, we consider each sentence in a *clinical case* as a separate input samples for our models. In a preliminary phase to our submission, we studied the effects of various hyperparameters and the generalization error of our models by splitting the original corpus of clinical cases in three parts: (1) a *training set* (60% of the original corpus) used to train the model, (2) a *validation set* (20% of the original corpus) to evaluate the effects of hyperparameters and (3) a *test set* (20% of the original corpus) to evaluate the ability of our models to generalize to unseen data. We fine-tune our models with a Google Colab environment, which provided us a Tesla T4 GPU.

## 5.2. Results

### NER hyperparameters and evaluation

We studied the effects of different hyperparameters on our validation set:

- `learning rate`: the initial learning rate for AdamW optimizer. Initialized to `5e-5`.
- `weight decay`: the weight decay to apply to all layers except all bias and LayerNorm weights in AdamW optimizer. Initialized to `0` (no weight decay applied)
- `batch size`: the batch size per device (e.g. CPU, GPU) for training. Initialized to `16`.

For each experiment, we train the NER model for two epochs — at the end of the selection process we will analyze the effects of an increased number of epochs. Our search for hyperparameters divides into two stages: in the first stage, we make hyperparameters vary in large ranges with the aim to detect a smaller range where we will perform a *grid search*. All the different configurations and associated performance results are listed in Table 1.

In the second stage, we perform a grid search based on a uniform distribution within the following hyperparameters ranges (which have been chosen based on the results of the first stage):

- `learning rate: [7e-5, 8e-5]`
- `weight decay: [0.1, 0.2]`
- `batch size: 8`

Given the best results from the grid search, we increased the number of training epochs with an *early stopping* criterion, by stopping training when the performance on the validation set does not increase for 5 consecutive epochs. The final preliminary results and the generalization error are shown in Table 2.

**Table 1**
NER hyperparameter selection (first stage)

| batch size | learning rate | weight decay | MiP | MiR | MiF1 |
|---|---|---|---|---|---|
| 16 | 5e-5 | 0.0 | 0.7199 | 0.7759 | 0.7521 |
| 16 | 4e-5 | 0.0 | 0.7136 | 0.7677 | 0.7448 |
| 16 | 3e-5 | 0.0 | 0.7095 | 0.7749 | 0.7460 |
| 16 | 2e-5 | 0.0 | 0.6805 | 0.7672 | 0.7263 |
| 16 | 1e-5 | 0.0 | 0.6209 | 0.7175 | 0.6704 |
| 16 | 6e-5 | 0.0 | 0.7274 | 0.7836 | 0.7598 |
| 16 | 7e-5 | 0.0 | 0.7370 | 0.7822 | 0.7624 |
| 16 | **8e-5** | 0.0 | **0.7400** | **0.7836** | **0.7642** |
| 16 | 9e-5 | 0.0 | 0.7331 | 0.7827 | 0.7571 |
| 16 | 1e-4 | 0.0 | 0.7373 | 0.7754 | 0.7612 |
| 16 | 8e-5 | 0.1 | 0.7375 | 0.7885 | 0.7675 |
| 16 | 8e-5 | **0.2** | **0.7428** | **0.7865** | **0.7694** |
| 16 | 8e-5 | 0.3 | 0.7396 | 0.7846 | 0.7668 |
| **8** | 8e-5 | 0.2 | **0.7479** | **0.7865** | **0.7722** |

**Table 2**
Final preliminary NER results

| epochs | batch size | learning rate | weight decay | MiP | MiR | MiF1 | |
|---|---|---|---|---|---|---|---|
| 18 | 8 | 8.516e-5 | 0.1844 | 0.7814 | 0.8031 | 0.7921 | best hyper-parameters |
| 18 | 8 | 8.516e-5 | 0.1844 | 0.7738 | 0.7931 | 0.7833 | internal test set error |

## EL evaluation

We evaluated results of our linking module with and without the *gazetteer*: challenge organizers declared that it contains all the possibile links to all the entity mentions in the test set. However, its size (113609 concepts) is much higher w.r.t. the number of concepts in our training set (2430 concepts). When a concept does not appear in the training set, its embedding is determined by its name and description, which could result in many "noisy" concepts leading to wrong linking results. Table 3 reports results on our "internal" test set (a 20% subset of the training files provided for entity linking) obtained with and without the gazetteer, i.e. we considered only concepts appearing at least one time in the training set. Since results are equivalent, we decided to keep the gazetteer for our submission.

**Table 3**
Entity linking "internal" test set results with and without using the gazetteer.

| System | MiP | MiR | MiF1 |
|---|---|---|---|
| With gazetteer | 0.7374 | 0.7374 | 0.7374 |
| Without gazetteer | 0.7374 | 0.7374 | 0.7374 |

**Table 4**
Official results of BioASQ DisTEMIST NER task. We show our result, the second-best result and median result (computed by considering just the best MiF1 score for each participant team).

| System | MiP | MiR | MiF1 |
|---|---|---|---|
| Ours | **0.7915** | **0.7629** | **0.7770** |
| Second-best participant | 0.7434 | 0.7483 | 0.7458 |
| Median | 0.7146 | 0.6736 | 0.6935 |

**Table 5**
Official results of BioASQ DisTEMIST linking task. We show our result, the best result and median result (computed by considering just the best MiF1 score for each participant team).

| System | MiP | MiR | MiF1 |
|---|---|---|---|
| Ours | 0.2814 | 0.2748 | 0.2780 |
| Best participant | **0.6207** | **0.5196** | **0.5657** |
| Median | 0.4795 | 0.2292 | 0.3102 |

## Leaderboard

Official results of the DisTEMIST track are reported in Table 4 (NER) and Table 5 (EL). Specifically, we show (1) our results, (2) results from the best participant team (second-best in case of NER, since our team is the first ranked), and (3) median results (computed by considering the best submissions of each participant team). While the domain-specific pre-training of the backbone network has been the key for a successful NER system, the EL solution seems to suffer from a design flaw. We can indeed observe a big discrepancy between results on our internal test set and the leaderboard, which may be caused by two main factors: (1) pipelined errors of NER and EL predictions and (2) the inappropriateness of the size of the training set: the gazetteer size (113609 concepts) suggests us that the leaderboard test set contains many concepts which are not present in our training set (2430 concepts). However, our context-based EL methodology computes embedded representations of concepts based on their occurrences in the training set, and all the other concepts are represented with their description provided with the gazetteer, which may be useless or even detrimental for similarity computation. Further investigations to handle the above-described problems are thus needed.

## 6. Conclusion

In this paper we have presented a simple but strong baseline based in biomedical Spanish language models. Specifically, we used a pre-trained biomedical Spanish transformer backbone network to fine-tune a NER model and to perform EL with an embedding similarity-based approach. Results on the official leaderboard of the DisTEMIST track at BioASQ 2022 challenge show that our NER approach largely surpasses the other participant baselines, while the EL approach has to be further investigated to improve its generalization ability over new clinical cases.

## References


[1] Z. Nasar, S. W. Jaffry, M. K. Malik, Named entity recognition and relation extraction: State-of-the-art, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3445965. doi:10.1145/3445965.

[2] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 Database issue (2004) D267–70.

[3] A. Miranda-Escalada, S. L.-L. Luis Gascó, 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, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2022.

[4] S. Zhang, N. Elhadad, Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts, Journal of biomedical informatics 46 6 (2013) 1088–98.

[5] S. Zhang, N. Elhadad, Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts, Journal of Biomedical Informatics 46 (2013) 1088–1098. doi:https://doi.org/10.1016/j.jbi.2013.08.004, special Section: Social Media Environments.

[6] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, 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, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270. URL: https://aclanthology.org/N16-1030. doi:10.18653/v1/N16-1030.

[7] S. Sahu, A. Anand, Recurrent neural network models for disease name recognition using domain invariant features, 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. 2216–2225. URL: https://aclanthology.org/P16-1209. doi:10.18653/v1/P16-1209.

[8] 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.

[9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. URL: https://arxiv.org/abs/2005.14165. doi:10.48550/ARXIV.2005.14165.

[10] 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.

[11] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. URL: https://aclanthology.org/W19-1909. doi:10.18653/v1/W19-1909.

[12] E. T. R. Schneider, J. V. A. de Souza, J. Knafou, L. E. S. e. Oliveira, J. Copara, Y. B. Gumiel, L. F. A. d. Oliveira, E. C. Paraiso, D. Teodoro, C. M. C. M. Barra, BioBERTpt - a Portuguese neural language model for clinical named entity recognition, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Online, 2020, pp. 65–72. URL: https://aclanthology.org/2020.clinicalnlp-1.7. doi:10.18653/v1/2020.clinicalnlp-1.7.

[13] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, M. Villegas, Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, 2021. URL: https://arxiv.org/abs/2109.03570. doi:10.48550/ARXIV.2109.03570.

[14] L. Ramshaw, M. Marcus, Text chunking using transformation-based learning, in: Third Workshop on Very Large Corpora, 1995. URL: https://aclanthology.org/W95-0107.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692. doi:10.48550/ARXIV.1907.11692.

[16] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M. P. Richardson, R. Stewart, A. D. Shah, W. K. Wong, Z. Ibrahim, J. T. Teo, R. J. Dobson, Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit, 2020. URL: https://arxiv.org/abs/2010.01165. doi:10.48550/ARXIV.2010.01165.

[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.