Spanish Pre-Trained Language Models for HealthCare Industry

Jalaj Harkawat[1,2] and Tejas Vaidhya[1,2]

[1] Indian Institute of Technology, Kharagpur, INDIA
[2] Equal Contribution by both Authors

Abstract. Currently transformer based model have shown high accuracy and good prediction on downstream tasks like Named Entity Recognition, Sentiment analysis etc. But the terminologies used in Healthcare sector such as names of different diseases, medicines and departments makes it difficult to predict with high accuracy. In this paper we are going to show a system for Named Entity tagging based on BETO (Spanish BERT). Experimental results have shown that our model gives better results than the current baseline of MEDDOPROF Shared task.

Keywords: BERT · NER · Healthcare Industry · Transformers · BART · BETO

1 Introduction

Natural Language Processing (NLP) is a rapidly expanding subject with several applications, and we are utilising it to get more insights into our existing dataset. We all know how important our occupations and employment status are to our identities. Occupations have a significant influence on one's physical and mental health, as well as their habits and lifestyle choices. For the prevention and control of the negative health impacts of our occupations, an entire medical specialty, occupational medicine, is required (workplace accidents, short and long-term effect of exposition to toxic substances and pathogens, work-related mental health issues such as overburden and stress). The COVID-19 epidemic has highlighted this impact, since many people in certain vocations have been disproportionately impacted (for instance, health professionals and other essential workers).
"Tools that automatically detect these sociodemographic factors can help researchers to better characterize multiple health aspects related to specific occupations. However, up until now these entities have mostly been ignored. The MEDDOPROF Shared Task [12] takes a more comprehensive look at occupations, also considering employment statuses and non-paid activities. [4]"

IberLEF 2021, September 2021, Málaga, Spain.

1.1 Background

We generate a lot of data as a result of continual technological development and a fast-paced environment, and with advancements in technology, particularly deep learning techniques used in Natural Language Processing (NLP), there has been substantial improvement in Named Entity Recognition. Long ShortTerm Memory (LSTM) [9] and Conditional Random Field (CRF) [11], for example, have significantly improved their performance in biological Named entity recognition (NER) [17] in recent years. In this paper we are introducing our system for Named Entity Recognition tagging on the MEDDOPROF Dataset. We will use BETO [6] which is a BERT [7] based model trained on big Spanish corpus.

Our Code and fine-tuned model is available at:
https://github.com/jharkawat/meddoprof_shared_task

2 Task Description and Dataset

MEDDOPROF (Medical Documents Profession Recognition) is a shared task organized within the IberLEF 2021 workshop that focuses on to develop automatic occupation detection systems for Spanish medical texts. It has three sub-tracks and Shared Task-1 Named MEDDOPROF-NER is a Named Entity Recognition task, requires automatically finding mentions of occupations and classifying each of them as a profession, an employment status or an activity. It can be described as a token-level classification task.

Our sentence with n number of words is defined as:

$$A = \{a_1, a_2, a_3, a_4, a_5, a_6.....a_n\} \tag{1}$$

Then it can be classified into a label set with m labels:

$$y = \{l_1, l_2, l_3, l_4, l_5, l_6.....l_m\} \tag{2}$$

Given Named-Entity of type YYY. If entities of type YYY are immediately next to each other, the first word of the second entity will be tagged B-YYY in order to show that it starts another entity and the entities inside $B-YYY$ will be represented as $I-YYY$. For example, the sentence "equipo de psiquiatría" have the following labels $\{B-PROFESION, I-PROFESION, I-PROFESION\}$.

2.1 Dataset

The MEDDOPROF corpus [8] is a collection of 1844 clinical cases annotated with professions and employment statuses from over 20 different specialities. After many rounds of quality control and annotation consistency analysis, the corpus was annotated by a team of linguists and clinical specialists who followed specifically developed annotation standards before annotating the whole dataset. Each clinical case will be stored as a separate file in the corpus, which will be delivered in plain text with UTF8 encoding.
Refer to Figure 1 for an example of the corpus' annotation [15]

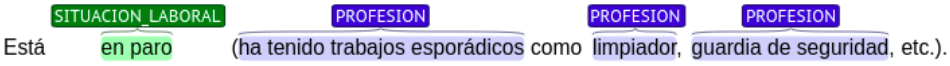

Fig. 1. BRAT annotation with profession, employment status labels. [3]

3 Approach

In part 3.1, we'll go over the BETO that we used in our final submission, then in section 3.2, we'll go over our problem-solving strategy, and finally, we'll go over which additional model we utilised during our tests.

Baseline We compare our system to the baseline provided by the organizers [1], which is a simple lookup system that uses the training set as a reference. It then examines if the extracted annotations are present in a fresh batch of text documents.

3.1 BETO

BETO is a BERT model trained on a big Spanish corpus [5] (ParaCrawl, EU-Bookshop [16], MultiUN [16], OpenSubtitles, DGC [16], DOGC [16], ECB, EMEA, Europarl, GlobalVoices [16], JRC, News-Commentary11 [16], TED, UN). BETO is around the same size (24-layer, 1024-hidden, 16-heads, 340M parameters) as a BERT-Base and was trained using the Whole Word Masking and Next Sentence Prediction classifiers. In most downstream tasks in the Spanish language, this surpassed the Best Multilingual BERT. Such language-specific bidirectional representations, we believe, are also important for our purpose.

3.2 Architecture

We first sub-word tokenize each token of sentences, using BETO's [6] wordpiece tokenizer from Huggingface [18] library and pass it through BETO Models BERT Transformer stacks (trained on big Spanish corpus) to extract contextualised domain specific representation. Then, for each word, we choose the representation of the first sub-word token and fine-tuning by training an additional feed-forward layer $\log(\text{softmax}(CW))$ that assigns the softmax probability distribution to each label.

The loss function used:

$$loss(x, class) = -\log\left(\frac{\exp(x[class])}{\sum_j \exp(x[j])}\right) = -x[class] + \log\left(\sum_j \exp(x[j])\right) \quad (3)$$

Additionally, we also tried using a Multilingual BERT (cased), which was pretrained model on the top 104 languages with the largest Wikipedia using a masked language modeling (MLM) objective. This model is case sensitive: it makes a difference between english and English.

4 Settings and Results

4.1 Experimental Settings

We keep maximum length of input sentence to 512 to consider long sentences. Large models (24-layer, 1024-hidden, 16-heads, 340M parameters) are trained for 4 epochs with batch size 16. We early stop the models using the valid set. The dropout probability was set to 0.1 for all layers. Optimization is done using Adam [10] with a learning rate of 5e-5. The remaining hyperparameters were kept same as BERT. We used the PyTorch [13] implementation of BERT from Huggingfaces tranformers library. An overview of these parameters is given in Table 1.

Table 1. Values of different parameters during the experiment

Parameter Name	**Value**
max length	512
learning rate	5e-5
weight decay	0.01
clip_grad	5
batch size	1
epoch number	20
min epoch number	5
patience	0.02
patience number	10

For selecting best models in experimental phase (i.e. before release of test set) we use split of 60/20/20 for train, dev and test respectively. For our final submission, we used a 70/30 split for train/valid set of initial data and a pre-trained BETO model. We also split sentences with more than 512 tokens to two or more sentences to get the desired model's input sentence length. To evaluate the performance of the system, an evaluation script [2] along with the dataset was provided by the organizers.

4.2 Results

The BETO models we propose is a competitive solution that performs better than the task's baseline. Tables 2 shows the system's performance on the test set,

while Table 3 contain our ablation study and result are based on partial matches, unlike the official results which uses exact matches. We also used multilingual BERT for performance comparison.
BETO performed better than other multi language model because it is trained on large domain specific(Spanish corpus) on the other hand multilingual BERT is trained on relatively less data with multiple languages.

Table 2. Results on test set

Models	**F1-score**	**Recall**	**Precision**
BETO	0.567	0.5	0.654
Baseline	0.486	0.508	0.465

Table 3. Performed Ablation study

Models(Partial match on provided data)	**F1-score**
BETO(mrm8488/BERT-spanish-cased-finetuned-ner)	0.753
Multilingual(BERT-base-multilingual-cased)	0.6342

5 Error Analysis

Domain specific pretrained transformer models have shown remarkable improvement in the majority of the downstream NLP tasks, but there are instances where BERT failed drastically. In this section we will try to find out some of the causes of failure in our system (BETO).

1. On our dataset, the BERT tokenizer is inefficient. Its lexicon does not include terminology from the healthcare industry, and it has not been trained in a language-specific setting. As a result, learning encoding based on improperly subtokenized words is difficult for the BERT model. Tokenizers could be trained on both biomedical and general text sets as a feasible approach.

2. The dataset contains a large number of phrases that lack Named Entity Recognition tags, resulting in a large number of negative entries and a poor F1-score. Increase the dataset size and eliminate the sentences with no or few NER tags available as a possible solution.
3. Because of small dataset size, our Transformer based model is not giving very good results. A Possible solution can be to add more positive datapoint and break the bigger sentences into smaller ones. This will also help with maintaining the token length less than 512 as desired by BERT based models.

6 Conclusion and Future Work

In this paper, we have presented a system based on BETO for the first sub-track of the MEDDOPROF Shared Task, held as part of the IberLEF 2021 workshop. We build our models keeping in mind the success of pre-trained models. It helps in generating bidirectional contextualized representation of each tokens that can be further utilised for task specific fine tuning.

As a future work, we would like to improve our current work by extending the work by performing layer by layer analysis of BERT and try to experiment with other Architectures like XLNet [19] and try to make more cost and memory efficient using adapter [14].

7 Acknowledgement

We would like to thank the organiser of Shared task MEDDOPROF for providing us this opportunity and to present our work.

References

1. Baseline code, `https://github.com/TeMU-BSC/meddoprof-baseline`
2. Evalutionscript, `https://github.com/TeMU-BSC/meddoprof-evaluation-library`
3. Example, `https://temu.bsc.es/meddoprof/data/`
4. Home page, `https://temu.bsc.es/meddoprof/`
5. Cañete, J.: Compilation of large spanish unannotated corpora (May 2019). https://doi.org/10.5281/zenodo.3247731, `https://doi.org/10.5281/zenodo.3247731`
6. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
8. Farré-Maduell, E., Lima-López, S., Miranda-Escalada, A., Briva-Iglesias, V., Krallinger, M.: Meddoprof corpus: test set (Jun 2021). https://doi.org/10.5281/zenodo.4889777, `https://doi.org/10.5281/zenodo.4889777`
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (Nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735, `https://doi.org/10.1162/neco.1997.9.8.1735`

10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
11. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. p. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
12. Lima-López, S., Farré-Maduell, E., Miranda-Escalada, A., Brivá-Iglesias, V., Krallinger, M.: Nlp applied to occupational health: Meddoprof shared task at iberlef 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. Procesamiento del Lenguaje Natural **67** (2021)
13. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
14. Rücklé, A., Geigle, G., Glockner, M., Beck, T., Pfeiffer, J., Reimers, N., et al.: Adapterdrop: On the efficiency of adapters in transformers (2020)
15. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: brat: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107. Association for Computational Linguistics, Avignon, France (Apr 2012), https://aclanthology.org/E12-2021
16. Tiedemann, J.: Parallel data, tools and interfaces in opus. In: Chair), N.C.C., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey (may 2012)
17. Vaidhya, T., Kaushal, A.: IITKGP at W-NUT 2020 shared task-1: Domain specific BERT representation for named entity recognition of lab protocol. In: Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020). pp. 268–272. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.wnut-1.34, https://www.aclweb.org/anthology/2020.wnut-1.34
18. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al.: Huggingface's transformers: State-of-the-art natural language processing (2020)
19. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding (2020)