NCU-IISR: Pre-trained Language Model for CANTEMIST Named Entity Recognition

Jen-Chieh Han, Richard Tzong-Han Tsai*

Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan

Abstract

Since BERT has brought a huge improvement in various NLP tasks, the great constructed pre-trained language model shows its power of being fine-tuned in other downstream tasks either. In this paper, NCU-IISR team adopted the Spanish BERT, BETO, as our pre-trained model, and the model was fine-tuned on CANTEMIST Named Entity Recognition (NER) data. Besides, we also compared it with another fine-tuned version, which was trained on an external Spanish medical text. Finally, our best score achieved an F1-measure of 0.85 in the official test set result for CANTEMIST-NER task.

Keywords

Electronic Health Records, Named Entity Recognition, Deep Learning, Pre-trained Language Model

1. Introduction

Biomedical text mining techniques can mine crucial information from literature in different languages, and the development in Spanish clinical data has become a novel research direction recently. Clinical texts contain immediate symptoms in reality. The demand is not only for those who are related to medical researchers, but the front of the hospital, such as doctors and patients. Because of the complexity of biomedical knowledge, what a key-point item we want to extract in electronic health records (EHRs) to analyze is setting up a base of the task tracking.

Cancer is the second-leading cause of death globally in 2018, and healthcare expenditures are growing for oncological treatments. The organizer of CANTEMIST (CANcer TExt Mining Shared Task) Track [1] is successful in previous shared tasks, and past experiments raise the new task. This year, they focus on named entity recognition (NER) at a critical type of concept related to cancer, that is to say, tumor morphology, and use a standard classification resource, International Classification of Diseases for Oncology (ICD-O) or eCIE-O-3.1 in Spanish, to code diagnosis. The Spanish language is the fourth-most spoken language in the world, and the language with these technologies have a potential of transfer learning to Italian, German, French, or even English. Furthermore, Spanish EHRs produce tens of thousands every 10 minutes, and therefore this task is collected in Spanish.

The CANTEMIST task involves three sub-tasks (CANTEMIST-NER, CANTEMIST-NORM, and CANTEMIST-CODING), and each has a particularly important use case scenario. For the

[*] Corresponding author.

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: joyhan@cc.ncu.edu.tw (J. Han); thtsai@csie.ncu.edu.tw (R.T. Tsai*)
ORCID: 0000-0003-0998-4539 (J. Han); 0000-0003-0513-107X (R.T. Tsai*)

CEUR Workshop Proceedings (CEUR-WS.org)

task summary, CANTEMIST explores the automatic assignment of eCIE-O-3.1 codes (Morfología neoplasia) to health-related documents in the Spanish language. In this project, we only participate in CANTEMIST-NER task, and the task requires finding automatically tumor morphology mentions, that is, the start and the end character offsets of tumor morphology in medical documents.

2. Pre-processing

The official dataset provides a set of text files with their corresponding annotation files. It includes 501 clinical cases in the training set, 500 documents in the dev(elopment) set, and 5231 documents in test and background set (only 300 documents test set are evaluated on the predictions). The dev set is divided into two subsets and each contains 250 documents. The second subset (dev-2) is released later than the first one (dev-1). We only used the dev-1 set because we did not have sufficient time to retrain our system on the whole dev set. Before we started training a model, we had to transform these input data into NER format. Thus, given a text with annotated name entities (NEs), we divided it into tokens and whether a token was in a NE (gold label).

First, we took GENIA Tagger's tokenizer[1], which was often used in biomedical tasks, to separate input text into tokens. If a token was in a NE, we gave it a label based on where it was located in NE positions. It is called IOB2 tag scheme [2]. Because we adopted BERT [3] method to train the NER model, and there was another tokenizer called WordPiece[2]. Therefore, an input token would be tokenized again and became word pieces with the vocabulary dictionary. Each label of tokens also expanded its numbers like word pieces and set the new one as the default label id. The final NER input format is shown in Table 1.

Table 1
Transforming text with NE (Bold) into NER input format from BERT. Example input text: **Carcinoma microcítico** de pulmón. There are label *B* (eginning), *I* (nside) and *O* (utside)

Token	Label
Car	*B*
##cino	None
##ma	None
micro	*I*
##cí	None
##tico	None
de	*O*
pulmón	*O*
.	*O*

[1] https://github.com/saffsd/geniatagger
[2] https://github.com/google-research/bert/blob/master/tokenization.py

3. Methods

In Cantemist-NER task, we submitted two prediction results, and the following subsection describes their methods.

3.1. Pre-trained Language Model

Bidirectional Encoder Representations from Transformers (BERT) [3] is a method of pre-training language representations and can be adapted to many natural language processing (NLP) tasks handily in fewer computational resources. BERT can use the same structure to be fine-tuned in various NLP tasks but also achieves state-of-the-art results.

Although BERT is written in English, they also provide Multilingual BERT, which contains over 100 languages. Because CANTEMIST-NER is a Spanish language dataset, Multilingual BERT[3] may not be sufficient. Then, we surveyed possible pre-trained language models for Spanish and found a Spanish BERT, called BETO [4] which has higher performances than Multilingual BERT, in Github. BETO is a BERT model and trained on over 300M lines of Spanish corpus such as Spanish Wikis. We adopted it as our pre-trained language model using a tool Transformers [5]. It adds a classification layer after BETO sequence output dropout. Cantemist-NER dev-1 result of BETO is shown in Table 2. We tuned the parameter *epoch* from 3 to 50 and chose the best F1 model as our first parameter setting of the submitted model.

Table 2
Cantemist-NER dev-1 results of BETO.

Epoch	Precision	Recall	F1
3	0.775	0.8	0.787
5	0.793	0.791	0.792
10	0.804	0.801	0.802
20	0.806	0.808	0.807
30	**0.816**	**0.819**	**0.818**
40	0.809	0.813	0.811
50	0.818	0.812	0.815

3.2. Spanish Biomedical Data

Because BETO is trained in the Spanish general domain, we thought that adding the last layer with clinical data before fine-tuning in NER may increase the final performance. Besides, we saw one team use the Spanish Health Corpus and got 90% F1 in last year [6]. Therefore, we found the resource and decided to take all Spanish text files from MedlinePlus in TEI format [7] instead of other XML file format. Then, we input raw texts over 37 thousand lines into BETO first. We called it BETO+Bio. Cantemist-NER dev-1 result of BETO+Bio is shown in Table 3. As the above model, we used the different *epoch* to train a model and found the top one as our second parameter setting of the submitted model.

[3] https://github.com/google-research/bert/blob/master/multilingual.md

Table 3
Cantemist-NER dev-1 results of BETO+Bio.

Epoch	Precision	Recall	F1
3	0.76	0.788	0.774
5	0.765	0.792	0.778
10	0.782	0.803	0.792
20	0.79	0.797	0.794
30	0.802	0.803	0.803
40	0.803	0.805	0.804
50	**0.81**	**0.802**	**0.806**

4. Results

We tried different hyperparameters to train two models on training and dev-1 sets. Our final Cantemist-NER test results are shown in Table 4. To our surprise, the best one is BETO and achieved an F1-measure of 0.85 which is slightly higher than BETO+Bio by 0.005. It may come from that the time limitation makes us unable to try different ways to transfer learning on the text of MedlinePlus.

Table 4
Cantemist-NER test results.

System	Precision	Recall	F1
BETO	0,849	0,851	**0.85**
BETO+Bio	0,84	0.85	0.845

5. Conclusion

Our NCU-IISR team constructed two models to predict Cantemist-NER labels, and both results are above 0.84 F1-measure. It showed that using the pre-trained language model can achieve high performance. We considered that adding a Spanish biomedical layer in the model would have a better result at first, however it got a slightly lower score. It might come from that the fine-tuned data was not enough, or we did not process text input properly before training. However, we believe that combining biomedical data before training in Cantemist-NER has the potential to help the model more fit in this task. Therefore, we would like to try possible ways in future work or find some errors to fix to improve results.

Acknowledgments

Appreciating Po-Ting Lai for giving us suggestions during the challenge and revising the working note.

References

[1] 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, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings, 2020.

[2] H.-J. Dai, P.-T. Lai, Y.-C. Chang, R. T.-H. Tsai, Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization, Journal of cheminformatics 7 (2015) S14.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[4] J. Cañete, G. Chaperon, R. Fuentes, J. Pérez, Spanish pre-trained bert model and evaluation data, in: to appear in PML4DC at ICLR 2020, 2020.

[5] 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. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Huggingface's transformers: State-of-the-art natural language processing, ArXiv abs/1910.03771 (2019).

[6] M. Stoeckel, W. Hemati, A. Mehler, When Specialization Helps: Using Pooled Contextualized Embeddings to Detect Chemical and Biomedical Entities in Spanish, in: Proceedings of the International Workshop on BioNLP Open Shared Tasks (BioNLP-OST), Association for Computational Linguistics SIGDAT and Asian Federation of Natural Language Processing, 2019. Accepted.

[7] M. Villegas, A. Intxaurrondo, A. Gonzalez-Agirre, M. Marimon, M. Krallinger, The mespen resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations, LREC MultilingualBIO: Multilingual Biomedical Text Processing (Malero M, Krallinger M, Gonzalez-Agirre A, eds.) (2018).