Clinical NER using Spanish BERT Embeddings

Ramya Vunikili[a,e], Supriya H N[b,e], Vasile George Marica[d,e] and Oladimeji Farri[a]

[a]Digital Technology & Innovation, Siemens Healthineers, NJ, USA
[b]Digital Technology & Innovation, Siemens Healthineers, Bangalore, India
[d]Siemens, Brasov, Romania
[e]All the authors have contributed equally.

Abstract

This paper presents an overview of transfer learning-based approach to the Named Entity Recognition (NER) sub-task from Cancer Text Mining Shared Task (CANTEMIST) conducted as a part of Iberian Languages Evaluation Forum (IberLEF) 2020. We explore the use of Bidirectional Encoder Representations from Transformers (BERT) based contextual embeddings trained on general domain Spanish text to extract tumor morphology from clinical reports written in Spanish. We achieve an F1 score of 73.4% on NER without using any feature engineered or rule-based approaches, and present our work as inspiration for further research on this task.

Keywords

Bidirectional Encoder Representations, BERT, NER, IberLEF 2020, Spanish embeddings, BETO, CANTEMIST

1. Introduction

There is a significant demand for automated analyses of electronic health record (EHR) documents to support clinical decision making and precision medicine. This is particularly true for documents written in Spanish language since nearly 10K of such documents are generated every 10 minutes in Spanish-speaking geographies [1].

According to the World Health Organisation (WHO), cancer was the second leading cause of death in 2018 [1]. Leveraging Natural Language Processing (NLP) techniques for cancer related EHR documents can not only expedite the decision making process but can also improve the quality of patient care by providing intrinsic information. Therefore CANTEMIST [1] focuses on automatic detection of the mentions related to tumor morphology through it's three independent tasks. We focus our work on the first sub-task, NER, by exploring contextual embeddings.

Contextualized language models rely heavily on large data sets to properly crystallize the deep embedding patterns specific to semantic meaning. As clinical text data on cancer reports is scarce, we chose to apply transfer learning using a BERT model [2], BETO [3], pre-trained on general domain Spanish text. Table 1 presents a comparison between the training corpus used for BETO and the CANTEMIST dataset.

Disclaimer: The concepts and information presented in this paper are based on research results that are not commercially available.

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: ramya.vunikili@siemens-healthineers.com (R. Vunikili); supriya.hn@siemens-healthineers.com (S.H. N); george.marica@siemens.com (V.G. Marica); oladimeji.farri@siemens-healthineers.com (O. Farri)
ORCID: 0000-0003-4629-3307 (R. Vunikili)

[1]https://www.who.int/news-room/fact-sheets/detail/cancer

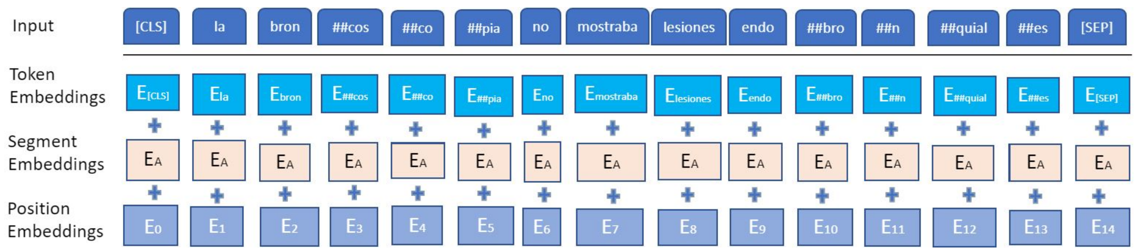

Figure 1: BETO embedding representation for the sentence: *la broncoscopia no mostraba lesiones endobronquiales.*

Table 1
BETO vs CANTEMIST corpus comparison

Criterion	BETO	CANTEMIST
Training corpus	ES Wiki; OPUS	-
Total number of tokens	3 billion	1.15 million
Unique tokens	31K	10.5 K

BETO has faithfully replicated the architecture behind the seminal contextualized embeddings inspired from Transformers [4] and is enhanced through training techniques like dynamic-masking [5] and whole-word-masking. As an example, Figure 1 shows the embedding of a Spanish sentence from the CANTEMIST corpus.

Also, since BETO has outperformed multilingual BERT (M-BERT) [2] on seven of the eight NLP tasks [3], we chose to use BETO as the base for the CANTEMIST NER task.

2. Related Work

Contextualized language models have provided improved performance for a myriad of NLP tasks by relying on a common deep network architecture. These models are often trained on a single large corpus of multilingual, general domain texts with subsequent fine-tuning on specific data sets through transfer learning.

One important reference in this field is the BERT language representation model which serves as basis for many zero-shot cross-lingual transfer. Trained on the top 104 Wikipedia versions, multilingual BERT has proven competitive in many NLP tasks. [6] Despite not benefiting from cross-lingual alignment, M-BERT outperforms models based on cross-lingual embeddings [7].

Such adaptability of M-BERT to various NLP tasks has been investigated end explained through the over-lapping effect of word-pieces across different languages. As such, common nouns, word roots, numbers, and URLs are mapped to a shared embedding space, determining co-occurring pieces [8]. Another study on the cross-lingual ability of BERT concludes that performance is relatively invariant with respect to word-pieces overlap or multi-head attention complexity[9] and suggests that the true versatility comes from a better network depth or a higher structural and semantic similarity between different languages.

Departing from the hypothesis that different languages have a common structural core to which M-BERT adapts during training, [10] follow the intuition of splitting a M-BERT sentence representation into a neutral (language agnostic) component and a specific language component. Through a series of tasks oriented towards language identification, language similarity, parallel sentence retrieval and word alignment, this study concludes that core cross-lingual representations are not neutral/general enough to mirror similar semantic structure. Consequently, multilingual embeddings are not good enough to solve difficult NLP tasks after zero-shot transfer learning.

In the same vein, an extensive study [11] regarding the internal structure of M-BERT used canonical correlation analysis [12] between similar representations in multiple languages. By looking at the similarity of deep layer representations, a divergence pattern was identified. M-BERT was not just mapping different languages into the same space but instead it was reflecting "linguistic and evolutionary relationships". Embeddings similarity was mostly identified in word-pieces rather than in word or character tokenization, with Romantic and Germanic languages clustered into different branches of the network.

A more targeted approach for transfer learning would be the identification of language families, where word-piece overlap, and similar grammar structure preserve the compact nature of a semantic representation. English to Spanish transfer learning for POS tagging has been shown improve performance when labeled data is scarce [13], or improve NER tasks when referring to proper nouns or niche concepts [14]. In the case where data is available in large quantities for individual languages, it is recommendable to combine specific language word representations with language-family models [15].

Considering these findings, we believe that multilingual contextualized embeddings are not optimal for those NLP tasks where either word-piece overlap, or semantic structure similarity are not high enough between pre-training corpus and task corpus. As such we have searched for a pre-trained BERT model that closely mimics the CanTeMiST data set. In ideal circumstances, such a model should have been pre-trained on Spanish EHR documents (labelled and/or unlabelled). However, we decided to explore the performance of the model trained on general domain Spanish text with fine-tuning, as the results can provide additional evidence to support the hypothesis that linguistic and evolutionary relationships can be learned from one domain and transferred to another.

3. Dataset and Experiments

We chose as task, the automatic named entity recognition of tumor morphology mentions in plain text medical documents.

The CanTeMiST dataset contains 6,933 de-identified clinical documents which are annotated for mentions related to tumor morphology, denoted by entity *MORPHOLOGIA_NEOPLASIA*, using the BRAT tool [16]. The annotations are done using well-established guidelines published by the Spanish Ministry of Health. Annotations have been made by clinical coding experts, according to eCIE-O-3.1 codes[2] following multiple iterations of quality control and annotation consistency. The choice of reports faithfully reflects the narative of electronic clinical reports. Table 2 summarises the data splits used as train, development and test sets along with the average number of tokens per report in each of these sets.

As a pre-processing step, all the reports are lower-cased and tokenized according to either sentences or sections of the reports so as to maintain a sequence length of less than or equal to 512. The sentence tokenizations are further broken-down to word-level tokens such that the start and end offsets of these tokens with respect to the original report are preserved. These word-level tokens are then encoded in BILOU format and given as input to fine-tune the BERT model on CANTEMIST dataset. During prediction time, all the tokens are O encoded as the ground truth is not provided. The output from

[2] https://eciemaps.mscbs.gob.es/ecieMaps/browser/index_o_3.html

Table 2
Summary of the data splits provided for CANTEMIST-NER sub-task.

Split	Dataset	Number of reports	Average number of tokens
Training Set	Train	501	739
Validation Set	Dev1	250	734
	Dev2	250	585
Testing Set	Test + Background	300 + 4932	348

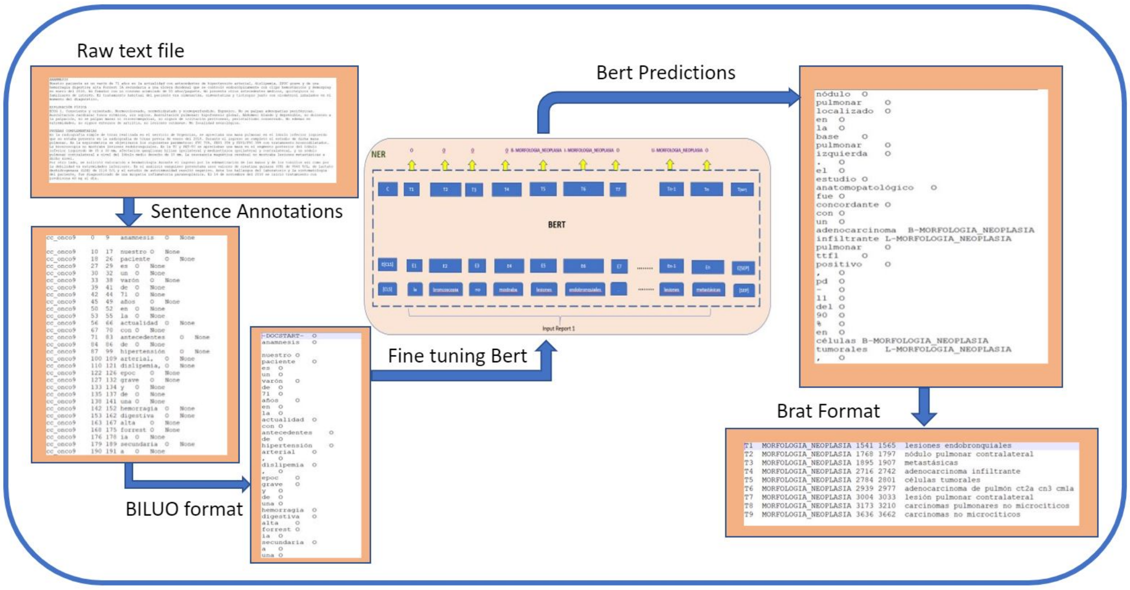

Figure 2: Overview of the prediction pipeline.

Table 3
Hyper-parameters of the BERT model

Parameter	Value
Learning rate	0.001
Optimizer	Adam
Maximum Sequence Length	512
Epochs	40

the BERT model is then gathered and post-processed to produce BRAT format. Figure 2 shows an overview of the pipeline used for prediction.

The BERT model is fine-tuned using AllenNLP platform [17] on NVIDIA Tesla V100 (32GB) GPU for 40 epochs, on the shuffled set composed of train, dev1 and dev2 data. Prediction is carried on both test and background sets. The hyper-parameters for the best model are summarised in Table 3.

Table 4
Performance metrics for NER.

Dataset	Precision	Recall	F1 Score
Test	72.7%	74.1%	73.4%

Figure 3: Excerpts from two reports along with named entities predicted by BERT. Green represents correctly identified mentions along with their spans. Yellow refers to mentions that are annotated to be a single entity but the model identified as separate entities. Red represents mentions that are not present in the ground truth but predicted by the model.

4. Results

Table 4 summarises the results obtained on test set using the official evaluation library for CanTeMiST [3] and Figure 3 presents excerpts from two reports and the entities predicted by the BERT model.

In order to account for the lower precision, it's worth studying the overlap between the vocabulary between BETO and CANTEMIST. The two vocabularies have an overlap of 24% which can be observed

[3]https://github.com/TeMU-BSC/cantemist-evaluation-library

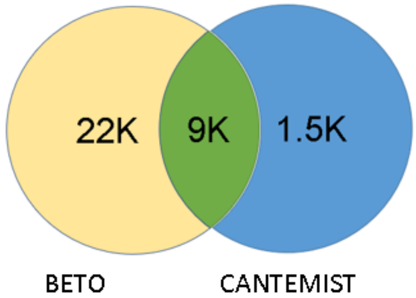

Figure 4: BERT and BETO vocabulary overlap

from Figure 4. Majority of these overlapped vocabulary contain suffixes such as '##s', '##l', '##al', '##a', '##op' that carry little-to-no information related to medical domain. And hence, the model struggled to differentiate between words such as *mycoplasma* (a bacteria) and *neoplasm* (abnormal growth of cells) which resulted in labelling the former as tumor related entity. In order to avoid such issues, it would be nice to add frequently occurring cancer related vocabulary to the unused tokens of BETO vocabulary so that the model can initialise different embedding irrespective of the suffix.

5. Future Work

As Spanish and English languages are syntactically similar, it might be safe to assume that some of the architectures that worked well for English might also translate to Spanish. One such model based on BERT and dynamic span graphs is DyGIEPP [18]. We plan on applying this architecture to CANTEMIST using the BETO embeddings as a next step.

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, 2020.

[2] 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).

[3] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: Practical ML for Developing Countries Workshop@ ICLR 2020, 2020.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[5] 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, arXiv preprint arXiv:1907.11692 (2019).

[6] S. Wu, M. Dredze, Beto, bentz, becas: The surprising cross-lingual effectiveness of bert, arXiv preprint arXiv:1904.09077 (2019).

[7] S. L. Smith, D. H. Turban, S. Hamblin, N. Y. Hammerla, Offline bilingual word vectors, orthogonal transformations and the inverted softmax, arXiv preprint arXiv:1702.03859 (2017).

[8] T. Pires, E. Schlinger, D. Garrette, How multilingual is Multilingual BERT?, arXiv preprint arXiv:1906.01502 (2019).

[9] K. Karthikeyan, Z. Wang, S. Mayhew, D. Roth, Cross-lingual ability of multilingual bert: An empirical study, in: International Conference on Learning Representations, 2019.
[10] J. Libovickỳ, R. Rosa, A. Fraser, How language-neutral is Multilingual BERT?, arXiv preprint arXiv:1911.03310 (2019).
[11] J. Singh, B. McCann, R. Socher, C. Xiong, Bert is not an interlingua and the bias of tokenization, in: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019), 2019, pp. 47–55.
[12] H. Hotelling, Relations between two sets of variates, in: Breakthroughs in statistics, Springer, 1992, pp. 162–190.
[13] Z. Yang, R. Salakhutdinov, W. W. Cohen, Transfer learning for sequence tagging with hierarchical recurrent networks, arXiv preprint arXiv:1703.06345 (2017).
[14] J. L. C. Zea, J. E. O. Luna, C. Thorne, G. Glavaš, Spanish NER with word representations and conditional random fields, in: Proceedings of the sixth named entity workshop, 2016, pp. 34–40.
[15] J.-K. Kim, Y.-B. Kim, R. Sarikaya, E. Fosler-Lussier, Cross-lingual transfer learning for pos tagging without cross-lingual resources, in: Proceedings of the 2017 conference on empirical methods in natural language processing, 2017, pp. 2832–2838.
[16] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, 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, Association for Computational Linguistics, Avignon, France, 2012, pp. 102–107. URL: https://www.aclweb.org/anthology/E12-2021.
[17] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, L. S. Zettlemoyer, Allennlp: A deep semantic natural language processing platform, 2017. arXiv:arXiv:1803.07640.
[18] D. Wadden, U. Wennberg, Y. Luan, H. Hajishirzi, Entity, Relation, and Event Extraction with Contextualized Span Representations, in: EMNLP/IJCNLP, 2019.