Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210184

Transfer Learning for Classifying Spanish and English Text by Clinical Specialties

Alexandra POMARES-QUIMBAYA [a,1], Pilar LÓPEZ-ÚBEDA [b], Stefan SCHULZ [c]
[a] *Pontificia Universidad Javeriana, Bogotá, Colombia*
[b] *Universidad de Jaén, Andalucía, Spain*
[c] *Medical University of Graz, Austria*

Abstract. Transfer learning has demonstrated its potential in natural language processing tasks, where models have been pre-trained on large corpora and then tuned to specific tasks. We applied pre-trained transfer models to a Spanish biomedical document classification task. The main goal is to analyze the performance of text classification by clinical specialties using state-of-the-art language models for Spanish, and compared them with the results using corresponding models in English and with the most important pre-trained model for the biomedical domain. The outcomes present interesting perspectives on the performance of language models that are pre-trained for a particular domain. In particular, we found that BioBERT achieved better results on Spanish texts translated into English than the general domain model in Spanish and the state-of-the-art multilingual model.

Keywords. Transfer learning, Classification, Natural Language Processing, Spanish

1. Background

The transfer of knowledge across several learning tasks in neural network based systems has already been described in 1996 [1], inspired by the idea that people intuitively use their previously learned experience to define and solve new problems. Over the past decade, transfer learning has firmly established itself as a machine learning (ML) approach, where a model trained for a source task is reused as the starting point to build a model for a new target task [2]. This approach is particularly useful for building models in contexts where training data are scarce or skewed. This is typical when applying ML to texts and images of limited size and accessibility, e.g. in clinical contexts [3].

Transfer learning has demonstrated its potential by its increased used in natural language processing (NLP) tasks, where language models have been pre-trained on large public corpora and then tuned to specific texts [4]. Like in most NLP research, English is the leading language, so that pre-trained language models in English like BERT [5] have first gained popularity by providing state-of-the-art solutions to tasks such as Named Entity Recognition, Relation Extraction and Question Answering. BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning

[1]Corresponding Author: Pontificia Universidad Javeriana, Cra. 7 No 40-62, Bogotá, Colombia; E-mail: pomares@javeriana.edu.co

on both left and right contexts in all layers. BERT uses a popular attention mechanism called transformer [6], which takes into account the context of words. Thus, pre-trained BERT models can be fine-tuned with just one additional output layer to create a multi-class classification model. This layer assigns a single code to a document. Soon after the launch of BERT, mBERT (Multilingual BERT) was pre-trained on the concatenation of monolingual Wikipedia corpora from 104 languages, showing good results for most of the European languages, particularly where languages share word order features [7].

One of the variations is BioBERT (BERT for Biomedical Text Mining), a model pre-trained on large-scale biomedical corpora [8], *viz.* PubMed abstracts and PubMedCentral full texts, outperforming BERT for biomedical text mining tasks. Shortly after BioBERT, a BERT variation pre-trained on a Spanish corpus gave rise to BETO, which is similar in size to BERT and trained using the Whole Word Masking technique [9].

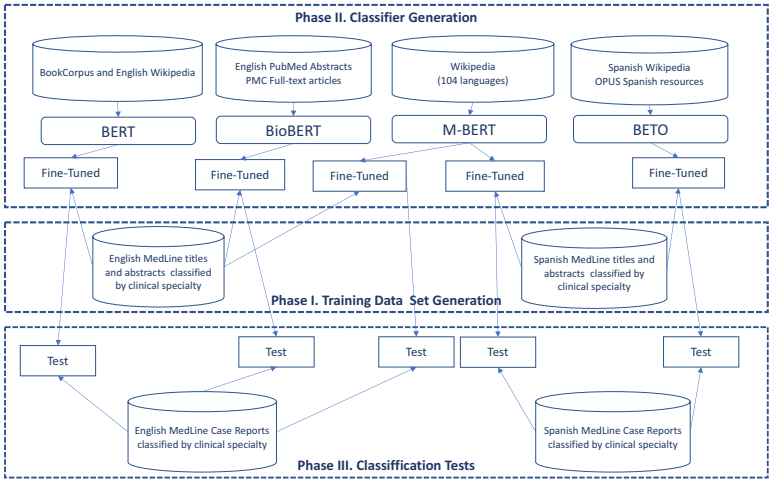
We describe the application of pre-trained transfer models to a Spanish biomedical document classification task using one or more labels that denote clinical specialties. Their recognition is important for downstream clinical NLP tasks, because every specialty has its own sublanguage, particularly characterized by short forms and elliptic expressions that would be highly ambiguous across all medicine, whereas their meaning within a given specialty is mostly univocal. In the absence of clinical documents we used publications of the type case report. The descriptions of real clinical cases they contain were manually extracted from full-text articles, after retrieving them from open access journals via PubMed. The main goal was to analyze how clinical specialty classification of Spanish texts performs using BETO and mBERT. Besides, we compared the results with BERT and BioBERT, after machine-translating the texts into English.

2. Method

Figure 1 depicts the design of our experiments. The first phase (I) involves the generation of the training data set, the second one (II) the classifiers applying four pre-trained models BERT, mBERT[2], and BioBERT [3] for the titles and abstracts in English; and BETO [4] and mBERT for the titles and abstracts in Spanish. The final phase (III) tests the models on the case reports extracted from MEDLINE [5]. The creation of the training data involves two automatic steps and a manual one. First, MEDLINE was searched for Spanish articles with freely accessible full texts, from which the type Case Report was excluded in order to avoid overlap with the test corpus. The query includes several filters to obtain specialty related papers, such as the journal name and the author affiliation (e.g. cardiology department). The clinical specialties we used for tagging were a subset of those included in MeSH [6]. From an initial set of 51 specialties, 18 particularly clinical ones (excluding dentistry) were selected, according to the typical dissection of health curricula and hospitals departments into classical *clinical* disciplines such as pediatrics, internal medicine, neurology, etc. This also involved the fusion of sub-specialties for which too few documents were available. Out of 194,527 PubMed records, the most common specialties were internal medicine, surgery, neurology and pediatrics.

[2] https://github.com/google-research/bert
[3] https://github.com/dmis-lab/biobert
[4] https://github.com/dccuchile/beto
[5] https://www.nlm.nih.gov/medline/medline_overview.html
[6] https://meshb.nlm.nih.gov

Figure 1. Comparison of Language pre-trained Models for Clinical Specialty Classification



The reuse of knowledge learned from previous tasks to a new task domain requires to fine-tune the pre-trained model. To this end, we chose from the following values: batch size, learning rate, max sequence and number of epochs. Table 1 illustrates the hyper-parameters and their tested options. The dataset was divided to carry out the hyper-parameter configuration; 80% was used for training and 20% for validation.

		Spanish		English		
Parameter	Options	BETO	mBERT	BioBERT	BERT	mBERT
Batch size	[8, 16, 32]	8	16	8	16	16
Max sequence	[256, 512]	512	512	512	512	512
Learning rate	[2e-5, 3e-5]	2e-5	3e-5	2e-5	2e-5	3e-5
Epoch	[3, 4, 5]	3	5	4	5	5

Table 1. Hyperparameters tested and options chosen in each model

The creation of the test data set involved four steps. Using the PubMed search interface, MEDLINE was searched for Spanish articles of the publication type case report, for which freely accessible full texts were available (Step 1). For each full text, a domain expert (last author) identified and copied the passages that described an individual clinical case (Step 2), obtaining, out of 227 articles, 263 case descriptions in Spanish. Each description was manually annotated with up to three clinical specialty tags (Step 3). The clinical specialties used for this tagging process were the same used for the creation of the training data set. The creation of a Spanish–English parallel corpus was done by translating the case descriptions into English (Step 4) using Google Translate[7]. The quality of translation was analyzed by the authors on a sample of 15% of the cases stratified by specialty and considered good. No manual correction of the translated texts were done. From the 263 available cases, the most common specialties were internal medicine, pediatrics and surgery. There was no case report for seven specialties. The label distribution of each dataset used with respect to the selected specialties is available on GiHub[8].

[7] https://pypi.org/project/googletrans/
[8] https://github.com/plubeda/mie_2021/blob/main/Distribution-of-labels.md

3. Results

The results employed the usual metrics in text classification, *viz.* precision, recall and F-score using the micro-average. Table 2 shows first the results of the Spanish corpus classified using BETO and mBERT. The language-specific model (BETO) achieved an F-score of 61.2%. The classification of the translated (English) corpus yielded the best results with the language and domain-specific model (BioBERT), reaching an F-score of 63.3%. Related to the multilingual pre-trained language model mBERT, in both languages we got the lowest results (55.61% F-score for English and 54.16% F-score for Spanish), being these results very similar, which is not surprising because for both languages mBERT uses the same information source for training (Wikipedia).

Language	System	Precision (%)	Recall (%)	F-score (%)
Spanish	BETO	54.61	69.59	61.20
	mBERT	58.76	50.23	54.16
English	BioBERT	66.25	60.60	63.30
	BERT	54.53	59.68	56.99
	mBERT	64.16	49.08	55.61

Table 2. Multi-label classification performance in clinical specialties for Spanish and English.

Furthermore, we have performed an additional evaluation that used the training dataset and dividing it into subsets in order to investigate which results we achieved with the type of texts used for training[9]. The results obtained are similar to those for case reports. However, this additional evaluation obtained higher values and BETO applied to the Spanish dataset achieved the best F-score value.

4. Discussion and Conclusion

The results show, in general, only moderate results, with no F-score greater than 63.3 %. This is explainable from the complexity of clinical cases, which often belonged to several clinical disciplines. The annotator had to take many boundary decisions. E.g., all cases of children and adolescents were annotated with "pediatrics", even if the disorder in focus belonged to neurology or surgery, for which additional tags had to be added. The case descriptions rarely clarified by which kinds of specialists patients were treated. Due to these contingencies, an inter-annotator agreement (IAA) would probably have yielded only fair agreement values. Clinical documents such as discharge summaries would have constituted a better gold standard, but, even here, one and the same clinical case could have been managed by different departments and specialists. E.g., a fracture can be treated by general surgeons, orthopedists or traumatologists, dependent on the institution. The lack of IAA analysis constitutes a limitation, but it is not assumed to affect the main research question *viz.* the comparison between language models.

[9] https://github.com/plubeda/mie_2021/blob/main/Additional_results.md

In this comparison, mBERT can be considered as a baseline. Its limitation to Wikipedia probably leaves important content out, and the advantage for English might be due the size of the English Wikipedia compared to the Spanish one. One could argue that particularly medical use cases could benefit from a highly multilingual language model, given the fact that many medical terms are similar across Western languages (e.g., *Appendicitis, Apendicite, Appendicite, Appendizitis, Apendicită*).

BETO's much higher performance surprises in comparison with BERT. Neither one is particularly trained on biomedical content, but the Spanish model outperforms the English one by four percentage points. That BioBERT outperforms BERT is not surprising because it was trained on the same text genre from which we took our examples. Our results demonstrate the value of BETO and raise the hypothesis that – alike BioBERT vs. BERT – there is still potential for optimization if a Spanish model is trained with medical content. That there is no large Spanish corpus comparable to PubMed and PMC might be mitigated by a massive use of machine translated texts, given our experiences with the good quality of Google translations. That machine-translation of Spanish texts offers a solution to the text classification problem, thus reducing the need for specific Spanish language models could be a conclusion of our study. To what extent this can be applied to clinical texts has to be further investigated. In any case, the training of ML models is a major desideratum in the clinical NLP community, although the exchange of clinical ML models between institutions is still fraught with complicated legal and ethical issues.

Acknowledgements

LIVING-LANG project [RTI2018-094653-B-C21] of the Spanish Government and the Fondo Europeo de Desarrollo Regional (FEDER) partially supported this work.

References

[1] Thrun S. Is learning the n-th thing any easier than learning the first? In: Advances in neural information processing systems; 1996. p. 640–646.

[2] Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering. 2010;22(10):1345–1359.

[3] Lv X, Guan Y, Deng B. Transfer learning based clinical concept extraction on data from multiple sources. Journal of Biomedical Informatics. 2014;52:55 – 64. Special Section: Methods in Clinical Research Informatics.

[4] Howard J, Ruder S. Universal language model fine-tuning for text classification. In: ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers). vol. 1; 2018. p. 328–339.

[5] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

[6] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems; 2017. p. 5998–6008.

[7] Pires T, Schlinger E, Garrette D. How Multilingual is Multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 4996–5001.

[8] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019 Sep.

[9] Cañete J, Chaperon G, Fuentes R, Ho JH, Kang H, Pérez J. Spanish Pre-Trained BERT Model and Evaluation Data. In: PML4DC at ICLR 2020; 2020. .