Proceedings of MultilingualBIO 2020, pages 29–32
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Transfer learning applied to text classification in Spanish radiological reports

Pilar López-Úbeda*, Manuel Carlos Díaz-Galiano*, L. Alfonso Ureña-López*, Maria-Teresa Martín-Valdivia*, Teodoro Martín-Noguerol†, Antonio Luna†
*Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{plubeda, mcdiaz, laurena, maite}@ujaen.es
†MRI Unit, Radiology Department, HT médica. Carmelo Torres 2, 23007 Jaén, Spain
{t.martin.f, aluna70}@htime.org

Abstract

Pre-trained text encoders have rapidly advanced the state-of-the-art on many Natural Language Processing tasks. This paper presents the use of transfer learning methods applied to the automatic detection of codes in radiological reports in Spanish. Assigning codes to a clinical document is a popular task in NLP and in the biomedical domain. These codes can be of two types: standard classifications (e.g. ICD-10) or specific to each clinic or hospital. In this study we show a system using specific radiology clinic codes. The dataset is composed of 208,167 radiology reports labeled with 89 different codes. The corpus has been evaluated with three methods using the BERT model applied to Spanish: Multilingual BERT, BETO and XLM. The results are interesting obtaining 70% of F1-score with a pre-trained multilingual model.

Keywords: Transfer Learning, BERT Model, Spanish Radiological Reports, CT Scanning

1. Introduction

Radiology reports are text records taken by radiologists that detail the interpretation of a certain imaging modality exam including a description of radiological findings that could be the answer to a specific clinical question (patient's symptoms, clinical signs or specific syndromes). Structured text information in image reports can be applied in many scenarios, including clinical decision support (Demner-Fushman et al., 2009), detection of critical reports (Hripcsak et al., 2002), labeling of medical images (Dreyer et al., 2005; Hassanpour et al., 2017; Yadav et al., 2013), among other.

Natural language processing (NLP) has shown promise in automating the classification of free narrative text. In the NLP area this process is named Automatic Text Classification techniques (ATC). ATC is an automated process of assigning set of predefined categories to plain text documents (Witten and Frank, 2002).

The health care system employs a large number of categorization and classification systems to assist data management for a variety of tasks, including patient care, record storage and retrieval, statistical analysis, insurance and billing (Crammer et al., 2007; Scheurwegs et al., 2017; Wang et al., 2016). One of these classification systems is the International Classification of Diseases, Ten Version (ICD-10[1]). In 2017 a challenge was born at CLEF where the aim of the task was to automatically assign ICD-10 codes to the text content of death certificates in different languages such as English, French (Névéol et al., 2017), Hungarian, Italian (Névéol et al., 2018) or German (Dörendahl et al., 2019).

Regarding ATC, many techniques have been applied and studied. In traditional machine learning the most common algorithms known in the radiology community are: Naive Bayes, decision trees, logistic regression and SVM (Wang and Summers, 2012; Wei et al., 2005; Perotte et al., 2014). On the other hand, Recurrent Neural Networks (RNN) are used for sequence learning, where both input and output are word and label sequences, respectively. There are several studies related to RNN using Long Short-Term Memory (LSTM) (Tutubalina and Miftahutdinov, 2017) or CNN with an attention layer (Mullenbach et al., 2018). Finally, researchers have shown the value of transfer learning — pre-training a neural network model on a known task and then performing fine-tuning — using the trained neural network as the basis of a new purpose-specific model. BERT model is one of the best known models nowadays. BERT has also been used for multi-class classification with ICD-10 (Amin et al., 2019) obtaining good results with minimal effort.

This study is in the initial phase and it is focuses on automatic code assignment in Spanish, so it can also be an automatic multi-class classification task. The main contributions of this paper can be summarized as follows:

- We analyse the performance of the three transfer learning architectures using the BERT models in Spanish: Multilingual BERT, BETO and XLM.
- We achieve encouraging results for a collection of Spanish radiological reports.
- We also investigate the fine-tuning parameters for BERT, including pre-process of long text, layerwise learning rate, batch sizes and number of epochs.

2. Medical collection

Dataset is composed of 208,167 anonymized Computed Tomography (CT) examinations. This clinical corpus has been provided by the HT médica. Each report contains relevant information such as: reason for consultation, information regarding the hospital where the CT scan was conducted, type of scan (contrast or non-contrast), and location of the scan (body part).

Each radiology report requires a unique code from the 89 available codes. These labels are assigned according to

[1] https://icd.who.int/browse10/2016/en

the area where the scan was performed, the type of contrast (contrast or non-contrast) and other clinical indications such as fractures, trauma, inflammation, tumors, and so on. Figure 1 shows the most common codes in the dataset and the number of documents in which each label appears. We can see that the TX4, TC4 and TX5 codes are the ones that appear most frequently in the corpus.

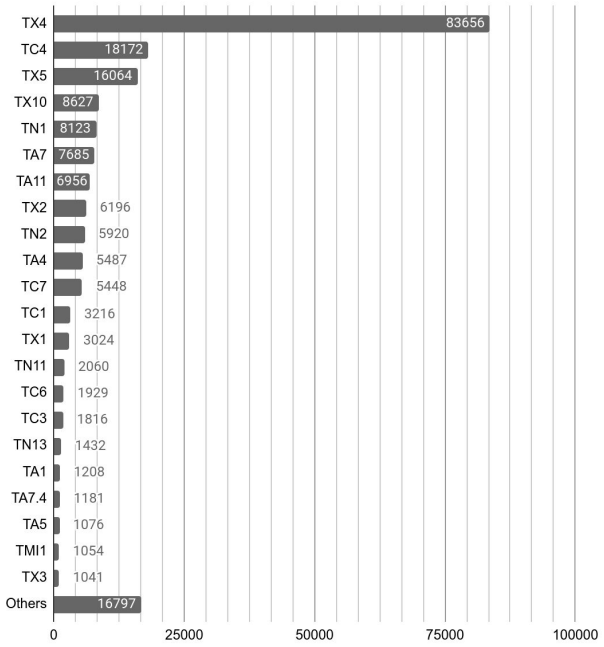

Figure 1: Most common labels and their frequency in the collection.

A weakness of the collection is that the text written by the specialists is in capital letters. Therefore, we pre-process the text by changing it to lower case.

Training, dev and test set The dataset was divided up to carry out the experimentation: 60% of the collection was used for the training set (124,899 documents), the development set was composed of 41,6434 documents (20%) and the remaining 20% for the test set (41,634 documents).
The sections of the CT examinations considered for this study were: the reason for the consultation, the location of the scan and the type of contrast used, avoiding hospital information because most of the examinations were done in the same hospital.

3. Code assignment methods

Transfer learning (Thrun, 1996) is an approach, by which a system can apply knowledge learned from previous tasks to a new task domain. This theory is inspired from the idea that people can intuitively use their previously learned experience to define and solve new problems.
For the automatic assignment of codes in Spanish, we have applied three transfer learning approaches based on BERT[2].
BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. BERT uses a popular attention mechanism called transformer (Vaswani et al., 2017) that takes into account the context of words.
As a result, the pre-trained BERT model 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.
In order to categorize radiology reports in Spanish, we have used three pre-trained models described below:

Multilingual (henceforth, M-BERT) follows the same model architecture and training procedure as BERT using data from Wikipedia in 104 languages (Pires et al., 2019). In M-BERT, the WordPiece modeling strategy allows the model to share embedding across languages.

BETO is a BERT model trained on a big Spanish corpus. BETO[3] is of size similar to a BERT for English and was trained with the Whole Word Masking technique (Cui et al., 2019).

XLM uses a pre-processing technique and a dual-language training mechanism with BERT in order to learn relations between words in different languages (Lample and Conneau, 2019). XLM presents a new training technique of BERT for multilingual classification tasks and the use of BERT as initialization of machine translation models.
In this study we show the performance of two XLM models: XLM trained with 17 languages (XLM-17) and trained with 100 languages (XLM-100)

4. Experiments and evaluation

4.1. Fine-tuning pre-trained parameters

In this step, we need to make decisions about the hyperparameters for the BERT model.
We use the BERT model with a hidden size of 768, 12 transformer blocks and 12 self-attention heads. For the optimizer, we leverage the *adam* optimizer which performs very well for NLP data and for BERT models in particular. For the purposes of fine-tuning, the authors recommend choosing from the following values: batch size, learning rate, max sequence and number of epoch. Table 4.1. illustrates the hyperparameters and their tested options, finally in each column we can see the model used and its selected parameter.

4.2. Results

In this section we present the results obtained by applying each BERT model. Since the corpus of radiological reports is in Spanish, we have applied the available models for this language in transfer learning.
The metrics used to carry out the experiments are the measures popularly known in the NLP community, namely macro-precision, macro-recall and macro-averaged F1-score.
Table 2 shows the results achieved and we can see that the results are encouraging, having a large list of codes to assign. XLM gets the best results by upgrading to BETO and

[2]https://github.com/google-research/bert

[3]https://github.com/dccuchile/beto

Parameter	Options	M-BERT	BETO	XLM-100	XLM-17
Batch size	[16, 32, 64]	32	16	16	16
Max sequence	[256, 512]	256	256	256	256
Learning rate	[2e-5, 3e-5]	3e-5	2e-5	2e-5	2e-5
Epoch	[3, 4, 5]	4	5	5	5

Table 1: Hyperparameters tested and options chosen in each model.

Pre-trainined Model	Precision	Recall	F1-score
M-BERT	65.41	62.07	62.33
BETO	69.86	65.34	66.34
XLM-100	**75.05**	69.10	70.64
XML-17	74.83	**69.79**	**70.84**

Table 2: Results obtained for code assignment in radiological reports.

M-BERT. XLM mixes several languages but it is enough to learn in the radiology reports and to detect the correct code. XLM-100 obtains the best precision (75%) and XLM-17 the best recall (69.7%). The best F1-score was also obtained with XLM-17 getting 70%.
Performing a brief analysis of the mislabeled codes, we found that the 23 worst-labeled codes had 2,443 documents to be trained, which is 1.96% of the total training set. In addition, the average number of training documents is 106, so they do not have enough information to learn. According to the evaluation of each code, Figure 2 shows the number of codes and their result ranges using the F1 score.

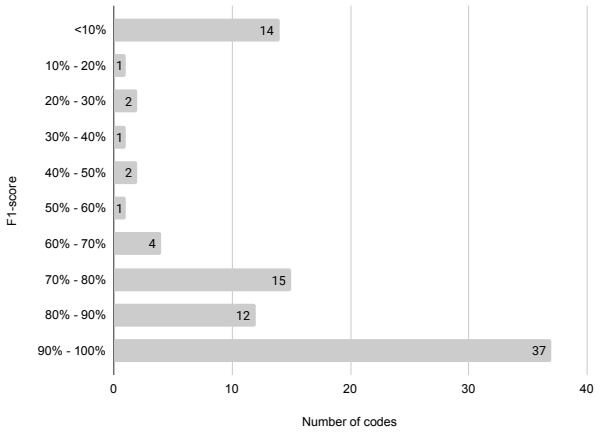

Figure 2: Results obtained in the F1-score and number of codes evaluated.

5. Limitations and future work

Our project is in a beginner's state and has limitations that need to be improved in the future. The limitations we found are shown below:

- Occasionally, the texts of the radiological reports are longer than allowed in the BERT model (max sequence of 512).
- The texts provided by the specialists are in capital letters, we pre-process the text by changing it to lower case.
- There are codes with few examples for training, so the system fails to classify.

We plan to make future improvements to the automatic classification system. These improvements can be summarized in the following points:

- We will perform a deep error analysis and see the behavior of each model applied to our corpus.
- We will analyze why XLM has achieved better results than BETO, being XLM trained for different languages.
- Strategies with embeddings to obtain the representation vector of each word will be used in future work.
- We will make changes to the model, for example, adding new layers or concatenating new features extracted from the corpus.
- We will improve BERT's vocabulary to find more words related to the biomedical domain. BioBERT (Lee et al., 2019) currently exists for English, we could make an adaptation or create a similar model with Spanish.
- There are parts of the text that are more important than others, for example the location of the exploration, in the future we plan to detect these features so that the model learns better.

6. Conclusion

In this study we conducted a multi-class task to detect codes in radiology reports written in Spanish. We have carried out experiments that are the state-of-the-art pre-training for NLP: BERT model. We apply different approaches using this model such as Multilingual BERT, BETO and XLM.
Recent advances in transfer learning model have opened another way to extract features and classify medical documents. We have a collection of over 200,000 CT scans and each text can have 89 possible codes. Each code is associated with the document for a reason. The most important reasons include: location of the body where the CT scan was performed or a previous finding or disease.
Using the XLM algorithm trained with 17 different languages we obtain a 70% of F1-score, detecting that the worst predictions are those codes that have scarce examples to train.

This study is at an early stage so we have described limitations and future work to further improve the code assignment task.

7. Acknowledgements

This work has been partially supported by the Fondo Europeo de Desarrollo Regional (FEDER), LIVING-LANG project (RTI2018-094653-B-C21), under the Spanish Government.

8. Bibliographical References

Amin, S., Neumann, G., Dunfield, K., Vechkaeva, A., Chapman, K. A., and Wixted, M. K. (2019). Mlt-dfki at clef ehealth 2019: Multi-label classification of icd-10 codes with bert. *CLEF (Working Notes)*.

Crammer, K., Dredze, M., Ganchev, K., Talukdar, P., and Carroll, S. (2007). Automatic code assignment to medical text. In *Biological, translational, and clinical language processing*, pages 129–136.

Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., and Hu, G. (2019). Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Demner-Fushman, D., Chapman, W. W., and McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.

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

Dörendahl, A., Leich, N., Hummel, B., Schönfelder, G., and Grune, B. (2019). Overview of the clef ehealth 2019 multilingual information extraction. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*.

Dreyer, K. J., Kalra, M. K., Maher, M. M., Hurier, A. M., Asfaw, B. A., Schultz, T., Halpern, E. F., and Thrall, J. H. (2005). Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. *Radiology*, 234(2):323–329.

Hassanpour, S., Langlotz, C. P., Amrhein, T. J., Befera, N. T., and Lungren, M. P. (2017). Performance of a machine learning classifier of knee mri reports in two large academic radiology practices: a tool to estimate diagnostic yield. *American Journal of Roentgenology*, 208(4):750–753.

Hripcsak, G., Austin, J. H., Alderson, P. O., and Friedman, C. (2002). Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology*, 224(1):157–163.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Névéol, A., Robert, A., Anderson, R., Cohen, K. B., Grouin, C., Lavergne, T., Rey, G., Rondet, C., and Zweigenbaum, P. (2017). Clef ehealth 2017 multilingual information extraction task overview: Icd10 coding of death certificates in english and french. In *CLEF (Working Notes)*.

Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikan, L., Ramadier, L., Rey, G., and Zweigenbaum, P. (2018). Clef ehealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian. In *CLEF (Working Notes)*.

Perotte, A., Pivovarov, R., Natarajan, K., Weiskopf, N., Wood, F., and Elhadad, N. (2014). Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Scheurwegs, E., Cule, B., Luyckx, K., Luyten, L., and Daelemans, W. (2017). Selecting relevant features from the electronic health record for clinical code prediction. *Journal of biomedical informatics*, 74:92–103.

Thrun, S. (1996). Is learning the n-th thing any easier than learning the first? In *Advances in neural information processing systems*, pages 640–646.

Tutubalina, E. and Miftahutdinov, Z. (2017). An encoder-decoder model for icd-10 coding of death certificates. *arXiv preprint arXiv:1712.01213*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, S. and Summers, R. M. (2012). Machine learning and radiology. *Medical image analysis*, 16(5):933–951.

Wang, S., Chang, X., Li, X., Long, G., Yao, L., and Sheng, Q. Z. (2016). Diagnosis code assignment using sparsity-based disease correlation embedding. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3191–3202.

Wei, L., Yang, Y., Nishikawa, R. M., and Jiang, Y. (2005). A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. *IEEE transactions on medical imaging*, 24(3):371–380.

Witten, I. H. and Frank, E. (2002). Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77.

Yadav, K., Sarioglu, E., Smith, M., and Choi, H.-A. (2013). Automated outcome classification of emergency department computed tomography imaging reports. *Academic Emergency Medicine*, 20(8):848–854.