CEUR-WS.org/Vol-2696/paper_135.pdf

MEDIA team at the CLEF-2020 Multilingual Information Extraction Task

Iker de la Iglesia[1], Mikel Martínez-Puente[2], Alexander Platas[1], Iria San Miguel[1], Aitziber Atutxa[1][0000−0003−4512−8633], and Koldo Gojenola[1][0000−0002−2116−6611]

[1] School of Engineering, Bilbao, University of the Basque Country (EHU/UPV),
{idelaiglesia004,platas51218,iria.san.miguel2000}@gmail.com,
{aitziber.atucha,koldo.gojenola}@ehu.eus
[2] Faculty of Science, University of the Basque Country (EHU/UPV)
mikelmpuente@gmail.com

Abstract. The aim of this paper is to present our approach (MEDIA) on the CLEF-2020 eHealth Task 1. The task consists in automatically assigning ICD10 codes (CIE-10, in Spanish) to clinical case documents, evaluating the prediction against manually generated ICD10 codifications. Our system took part in two different subtasks: one corresponding to Diagnosis Coding (CodiEsp-D) and the other to Procedure Coding (CodiEsp-P).

We approached the coding task as a two step system; a first step consisting of carrying out the named entity recognition (diagnoses and procedures) and a second step for assigning the right ICD10 code to the given entity (diagnosis or procedure). For the first step, namely the medical entity recognition, we employed a transfer learning strategy over pretrained Language Models by tuning them to the Named Entity Recognition task. The second step was dealt with edit distance techniques. We achieved our best results combining static and contextual word embeddings of Wikipedia and Electronic Health Records (∼100M words), with a Mean Average Precision (MAP) of 0.488 and 0.442 for diagnoses and procedures, respectively.

Keywords: Neural Networks · Levenshtein Distance · ICD Coding.

1 Introduction

Automatic clinical coding has received great attention and several systems and shared tasks have been organized in the last years, using knowledge-based and machine learning techniques. CodiEsp: Clinical Case Coding in Spanish Shared Task (eHealth CLEF 2020 – Multilingual Information Extraction) [1] is devoted to the automatic coding of clinical cases in Spanish, as part of the CLEF eHealth

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.

series [2]. Participant systems have to automatically assign ICD10 codes to clinical case documents evaluating the result against manually generated ICD10 codifications. This task will serve to generate new clinical coding tools for other languages and data collections.

Our system took part in two different subtasks: one corresponding to the detection and coding of diagnoses (CodiEsp-D) and the other to procedure coding (CodiEsp-P). We made use of different techniques, ranging from edit distance measures to neural approaches using the most recent architectures, including different types of embeddings taken from medical texts.

2 Related Work

The SemEval 2014 Task 7 [5] can be cited as an antecedent to the present competition, except for the number and types of entities to be identified (diseases and others), and the type of concept indexing (SNOMED-CT, compared to ICD10 in this work). Task 7 in SemEval 2014 comprised two subtasks, medical entity recognition and concept indexation. To tackle the first subtask, different teams used approaches as MaxEnt, SVM or CRF in combination with the extraction of syntactic and semantic attributes. The authors in [6] obtained the best results in strict F-Score with 78.5 on the development set and 81.3 on the test set. For the second subtask, namely Concept Indexation, the solutions proposed were very similar among the different teams. As in the NER task, the winner was [6] with an accuracy of 74.1 on the test set. Their solution was based on the cosine similarity using a Vector Space Model (VSM). Other teams also proposed a method based on edit distance, more precisely Levenshtein distance [4].

In SemEval 2015 (task 14) [7], the methods used were analogous to those used in SemEval-2014. In the best system, a CRF was used to detect entities and a SVM classifier to determine if these were joined or not (and thus catch discontinuous entities). Regarding Concept Indexing, they used basically customized look-ups, like Dictionary look-up (exact match of entity word permutations, LVG), Customized Dictionary look-up (split UMLS entities by function words), and Customized Dictionary look-up (list of possible UMLS spans and application of Levenshtein distance).

Recently, the PharmacoNER competition [9] proposed a similar task, but with the aim of identifying chemical, drug, and gene/protein mentions for clinical case studies written in Spanish. The evaluation of the task was divided in two scenarios: one corresponding to the detection of named entities and one corresponding to the indexation of named entities. Besides these competitions, improvements have been made mostly in the entity recognition subtask using neural networks such as Bi-LSTM + CRFs [8].

3 Resources and Methods

3.1 Resources

The CodiEsp corpus contains 3,751 (including background set) clinical cases in Spanish language, with 1,000 of them manually coded. The annotated corpus has been randomly sampled into three subsets: the train, the development, and the test set. The train set contains 500 clinical cases, and the development and test set 250 clinical cases each. The train and development cases were provided along with their corresponding annotations. The final collection of 1,000 clinical cases that make up the corpus had a total of 16,504 sentences, with an average of 16.5 sentences per clinical case. It contains a total of 396,988 words, with an average of 396.2 words per clinical case.

In addition to this corpus, we made use of medical domain related synonym dictionaries and word, character and contextual embeddings calculated from a medical corpus other than the one provided by the organizers.

3.2 Methods

The first experiment, applied to both diagnoses and procedures, consisted in finding word sequences in the text keeping a given edit distance proportional to the length of the sequence with respect to some ICD10 dictionary term (EditDistance, see below). Additionally, for the procedures, we made a similar experiment (UnorderedMatching) but, in this case, instead of matching the sentence's words in the exact order, we focused just on the words individually. If an entity's words matched all words of an entry from the dictionary, even with a different word order, it was interpreted as the same term.

In these two experiments, we made use of sliding windows. Initially the window covers the maximum size available up to a word limit (in this case 7 words). Then it tries to match the sentence with a known entity with one of the two methods above. If no match was found, and as long as the window's size was longer than 1 word, the window size decreases by one and tries to match the new shorter entity. When the window size is 1 and there is no matching entity found, the window will move one word forward. Finally, if a match is found, the window resets its size and moves forward to the word after the last word of the previous sequence.

With this approach, the entity recognition and the ICD code assignment was done in one unique step. For the rest of the experiments we pursue the task in two steps, first detecting the entities using transfer learning by tuning a pretrained LM and then applying edit distance.

Named Entity Recognition. Recently, the use of pretrained Language Models (LM) on huge amounts of data (ElMo [13], BERT [14], FLAIR [3]) has shown to obtain very good results in different tasks [10, 13]. Language Models are able to capture the distribution of a language by learning a probability distribution over word sequences. This has proven to be a good approach in sequence-to-sequence tasks like named entity recognition (NER) [12], especially when using bidirectional Language Models (BiLM) learned on both left-to-right and right-to-left directions [13]. From the different available Transfer Learning language Models we decided to use Flair [3]. This system, compared to the other mentioned alternatives, is computationally less expensive without harming the performance.

For this work, we trained two different LMs. We trained the first LM on 331,468 Electronic Health Records (EHR) containing ∼100M words from different Spanish hospitals (SpaEHR). The second one was trained on the Spanish Clinical Case Corpus (SPACCC[3]), a collection of 1,000 clinical cases from SciELO (Scientific Electronic Library Online). Clinical cases have the peculiarity of being a biomedical and medical literature version of EHRs, and therefore the language employed is mostly standard as opposed to conventional EHRs.

For the diagnosis NER subtask, we run four different experiments. The first one used edit distance as explained at the beginning of this section, where the entity recognition and the code assignment where done in one unique step. For the three other experiments we employed Flair [3], performing the NER task by means of a BiLSTM + CRF. In this case we used the same static word representation in each experiment and character based embeddings as well but fine-tuning the LMs we mentioned before. As static word representations, we employed FastText and word2vec embeddings learned from wikipedia and skip-Ngram embeddings learned from the EHRs corpus (SpaEHR). Part of this corpus was originally tagged with diagnoses so we fine-tuned the LM learned on the SpaEHR corpus adapting it to the diagnoses NER subtask. Therefore in this experiment (from now on SpaEHR-LM) we did not use the train, development and test sets provided by the organizers to fine-tune the LM. Noteworthy, the SpaEHR corpus is not tagged with discontinuous entities (in this case diagnoses). As a consequence such kind of entities cannot be recognized in the SpaEHR experiment. For the second run, although we used the same static embeddings (fasttext and word2vec from wikipedia and SkipNGram from SpaEHR), the LM was the one learned on the SPACCC corpus (from now on SPACCC-LM), that is to say, on the corpus provided by the organizers which contains more standardized clinical cases with tagged diagnoses and procedures. In this case, the training corpus contains discontinuous entities so we pre-processed the corpus to convert it into a tabular format with IBOES tags. And finally, the third experiment consisted in joining the entities found by the SpaEHR-LM and those found by the SPACCC-LM (Joint-LM).

It is important to mention that we employed the same SPACCC-LM model for diagnoses and procedures, since the SPACCC corpus provided by the organizer contained both diagnoses and procedures.

Edit distance. The normalization of given named entities consists in linking named entities to concepts in standardized medical terminologies, allowing

[3] https://doi.org/10.5281/zenodo.2560316

generalization across contexts. The task consists in assigning, to each term, its corresponding Concept Unique Index. For example, “fiebre”, “hipertermia” and “sindrome febril” are all normalized to the same ICD-10 code (r50.9). In our work, we made use of a Text Similarity based mapping from the given terms to different sets:

- The terms present in the training set. This set is limited but gives an account of standard and non-standard terms present in spontaneously written health records. These terms are a source of spontaneously written data, similar to those present in the test set. However, this set only covers a small fraction of the whole set of ICD-10 codes.
- ICD10 standard terms. This can be viewed as a dictionary covering all terms. However, the description is far from the terms found in spontaneous clinical cases.

A lookup table was built by traversing the training data, recording every entity and its corresponding ICD code, and directly applied on the test set. We tried to approximate the search to guarantee a matching, by using the Levenshtein distance [4], a method that quantifies the minimum number of operations required to transform one string into another using insertions, deletions or substitutions as the basic edit operations. We compute the distance between the input string and the set of terms taken as reference. We must take into account that the methods just try to match the chosen strings with terms of a dictionary of expressions or the list of entities present in the training set, ignoring the context around the entities. In order to improve the number of different entities that can be found, we included a synonym dictionary. Therefore, if no match was found the first time assigning a code to an entity, synonyms were applied to it, composing new candidate terms.

4 Results and Discussion

In this section we will present the results we have achieved for both subtasks (CodiEsp-D and CodiEsp-P) of Task 1. Multilingual Information Extraction. For this purpose we have compiled all the results in Table 1 and Table 2 respectively, where we can observe the results and the approach of each and every run we have made.

In the first subtask, i.e., with regard to diagnoses, we have submitted four different runs: *“SPACCC-LM”,“SpaEHR-LM”, “Joint-LM”, and “EditDistance”*. In Table 1 we can observe the different results according to the official metrics, i.e., MAP (Mean average precision for a set of queries is the mean of the average precision scores for each query) and other computed metrics such as MAP30, Precision (guessed codes/all of our predictions), Recall (guessed codes/all codes) and F-score (harmonic mean of Precision and Recall).

Overall, the best run is the one corresponding to *“Joint-LM”*, taking into account all the metrics. However, if we analyze each run and each metric one

Table 1. Results of the different runs of CodiEsp-D.

Run	MAP	MAP30	Precision	Recall	F-score
SPACCC-LM	0.457	0.457	**0.735**	0.543	0.625
SpaEHR-LM	0.405	0.405	0.633	0.518	0.570
Joint-LM	**0.488**	**0.487**	0.637	0.620	**0.629**
EditDistance	0.462	0.461	0.526	**0.630**	0.574

by one we can find some interesting information. For example, the Precision of *"SPACCC-LM"* is 0.735, which is much higher than the rest, whereas the Recall is 0.543, lower than others. We can also state that the worst run is *"SpaEHR-LM"*, because its figures are the lowest ones in almost all the metrics.

In the second subtask, namely procedures, we have submitted three different runs, *"EditDistance", "SPACCC-LM", and "UnorderedMatching"*. In Table 2 we can observe the different results according to official metrics.

Table 2. Results of the different runs of CodiEsp-P.

Run	MAP	MAP10	Precision	Recall	F-score
EditDistance	0.386	0.383	0.455	**0.520**	0.485
SPACCC-LM	**0.442**	**0.442**	**0.601**	0.412	0.489
UnorderedMatching	0.404	0.402	0.501	0.503	**0.502**

If we analyze each run and each metric we can observe that there are differences among them depending on which metric we use. For example, the Precision of *"SPACCC-LM"* is the highest one (0.601) while the Recall is lower than the other runs (0.412). In some other cases, such as in *"UnorderedMatching"*, both Precision and Recall remain almost constant: 0.501 and 0.503, respectively.

5 Conclusion and Future Work

The purpose of this work was to evaluate the feasibility of different approaches to medical entity detection and concept indexing using the International Classification of Diseases, ICD10. Entity detection was dealt with a sequential tagger that used word embeddings and contextual string embeddings acquired from Electronic Health Records (EHR), Clinical Cases and Wikipedia. Concept normalization was approached by Text Similarity techniques. The Levenshtein-based system obtained relatively good results, compared to neural network approaches, and this aspect deserves a further study of the strengths and weaknesses of each approach.

Acknowledgements

This work has been partially funded by the Spanish ministry (projects PROSAMED: TIN2016-77820-C3-1-R, DOTT-HEALTH: PID2019-106942RB-C31). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

1. Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estapé, J., Krallinger, M.: Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of CLEF eHealth 2020. Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings, (2020)
2. Goeuriot, L., Suominen, H., Kelly, L., Miranda-Escalada, A., Krallinger, M., Liu, Z., Pasi, G., Saez Gonzales, G., Viviani, M., Xu, C. Overview of the CLEF eHealth Evaluation Lab 2020. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S. Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L. and Ferro, N.(eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020). LNCS Volume number: 12260 (2020)
3. Akbik, A., Blythe, D., Vollgraf, R. Contextual string embeddings for sequence labeling. Proceedings of the 27th International Conference on Computational Linguistics, pages 1638–1649, (2018)
4. Levenshtein, V. Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics doklady, volume 10, number8, pages 707-710, (1966)
5. Pradhan, S., Elhadad, N., Chapman, W., Manandhar, S., Savova, G. SemEval-2014 Task 7: Analysis of Clinical Text. SemEval Workshop (COLING), pages 54-62, (2014)
6. Tang, Y., Zhang, J., Wang, B., Jiang, Y., Xu, Y. UTH_CCB: a report for semeval 2014–task 7 analysis of clinical text. SemEval Workshop (COLING), page 802, (2014)
7. Pathak, P., Patel, P., Panchal, V., Soni, S., Dani, D., Patel, A., Choudhary, N. ezDI: A Supervised NLP System for Clinical Narrative Analysis. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, pages 412-416, (2015)
8. Lample, G., Ballesteros, M., Subramanian, S., Subramanian, K., Dyer, C. Neural architectures for named entity recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 260-270, (2016)
9. Gonzalez, A., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., Krallinger, M. PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. Proceedings of the BioNLP Open Shared Tasks (BioNLP-OST), Association for Computational Linguistics (2019)
10. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. Improving Language Understanding by Generative Pre-Training.(2018)

11. Peters, M.,Neumann, M., Iyyer, M., Gardner, M., Clark, C. Lee, K., Zettlemoyer, L. Deep Contextualized Word Representations, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, Association for Computational Linguistics (2018)
12. Shreyas, S., Daniel Jr, R. BioFLAIR: Pretrained Pooled Contextualized Embeddings for Biomedical Sequence Labeling Tasks(2019),1908.05760 arXiv
13. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. Deep Contextualized Word Representations, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1,(2018)
14. Devlin, J., Chang, M., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, (2018), 1810.04805 arXiv