CEUR-WS.org/Vol-2696/paper_198.pdf

IAM at CLEF EHEALTH 2020: CONCEPT ANNOTATION in SPANISH ELECTRONIC HEALTH RECORDS

Sébastien Cossin[1,2][0000−0002−3845−8127] and Vianney Jouhet[1,2][0000−0001−5272−2265]

[1] Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, team ERIAS, UMR 1219, F-33000 Bordeaux, France
[2] CHU de Bordeaux, Pôle de santé publique, Service d'information médicale, Informatique et Archivistique Médicales (IAM), F-33000 Bordeaux, France
sebastien.cossin@u-bordeaux.fr

Abstract. In this paper, we describe the approach and the results of our participation in task 1 (multilingual information extraction) of the CLEF eHealth 2020 challenge. We tackled the task of automatically assigning ICD-10 diagnosis and procedure codes to Spanish electronic health records. We used a dictionary-based approach using only materials provided by the task organizers. The training set consisted of 750 clinical cases annotated by a medical expert. Our system achieved an F1-score of 0.69 for the detection of diagnoses and 0.52 for the detection of procedures on a test set of 250 clinical cases.

Keywords: Semantic annotation · Entity recognition · Natural Language Processing · Electronic Health Records · Spanish

1 Introduction

An electronic health record is a patient-centered record that contains medical information about a patient's medical history, past and current medications, lab results, diagnoses etc. Most of this medical information is provided by health care professionals in free text format. Free text has many advantages like familiarity, ease of use and freedom to express complex things [3]. However unstructured data are difficult to reuse and query to retrieve information. Natural Language Processing (NLP) develops methods to manage free-text data and extract information required by applications such as clinical decision support systems. A frequent step in a NLP pipeline is the detection of medical entities (treatment, diagnosis) with named entity recognition (NER) algorithms. Linking each detected entity to a terminology or ontology is essential to leverage the power of knowledge graphs that bring external knowledge and meaning [4,5].

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.

The objective of shared tasks is to foster the development of NLP tools. For many years CLEF eHealth has proposed challenges to solve several real-world problems of information extraction in free-text data.

In this paper, we describe our approach and present the results for our participation in the task 1 (multilingual information extraction) of the CLEF eHealth 2020 challenge [2,6]. This task focused on diagnosis and procedure coding of Spanish electronic health records. We addressed 3 subtasks:

1. ICD10-CM codes assignment. In this sub-track the systems must detect symptoms and diseases mentioned in clinical notes by predicting ICD10-CM codes (International Classification of Diseases, Tenth Revision, Clinical Modification) for each document.
2. ICD10-PCS codes assignment. In this sub-track the systems must detect procedures mentioned in clinical notes by predicting ICD10-PCS codes (International Classification of Diseases, Tenth Revision, Procedure Coding System) for each document.
3. Explainable AI. In this sub-track the systems must provide text annotations for each ICD10-CM and ICD10-PCS code prediction of sub-track 1 and 2.

We developed a biomedical semantic annotation tool for our own needs at Bordeaux hospital. The main motivation to participate in the challenge was to compare our system and to learn from others on a shared task.

2 Methods

In the following subsections, we describe the corpora, the terminologies used in this challenge, an exploratory analysis of the data and our system.

2.1 Corpora

The dataset provided by the organizers was called the CodiEsp corpus. The corpus comprised 1,000 Spanish clinical case studies selected by a practicing physician. Each clinical case was a plain text file and the filename was the identifier. The train set and the development set contained 500 and 250 clinical cases, respectively. For these sets, annotations were published.

3,001 clinical cases had to be annotated by the participants, of which 250 formed the test set and were only known by the organizers to avoid manual corrections.

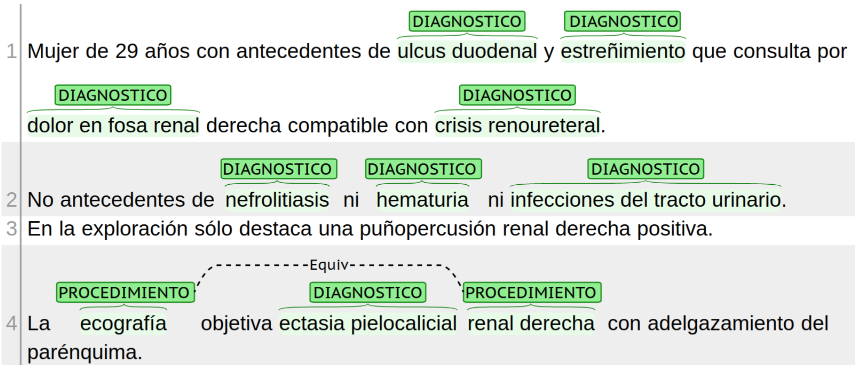
The annotation format was a tab-separated file with 2 fields for subtasks 1 and 2 corresponding to the clinical case identifier and an ICD10-CM or ICD10-PCS code. Three more fields were expected for the third substask: the start and end offset of each term detected and whether the code came from the ICD10-CM or the ICD10-PCS terminology.

2.2 Coding terminologies

Spanish versions of the ICD10-CM and the ICD10-PCS terminologies were provided by the organizers. The terminologies contained 98,288 and 75,789 different codes, respectively. In this task only 2,921 distinct codes (1.7%) were present in the train and development sets. Therefore, a vast majority of codes were not used while others were frequent.

2.3 Corpora exploration

The Brat annotation tool [7] was used to visualize the annotations made by the medical expert. To do so, a script was developed to transform the task file format to the Brat file format. Figure 1 presents a screenshot of the Brat interface with the first 4 lines of a clinical case from the development set.



1 Mujer de 29 años con antecedentes de ulcus duodenal y estreñimiento que consulta por dolor en fosa renal derecha compatible con crisis renoureteral.

2 No antecedentes de nefrolitiasis ni hematuria ni infecciones del tracto urinario.

3 En la exploración sólo destaca una puñopercusión renal derecha positiva.

4 La ecografía objetiva ectasia pielocalicial renal derecha con adelgazamiento del parénquima.

Fig. 1. The Brat interface was used to visualize the annotations made by the medical expert. Each annotation was linked to an ICD10-CM or ICD10-PCS code (not shown here)

Four key insights emerged from this visualization:

- All clinical diagnoses and procedures detected in a clinical note had to be coded. It's different from medical coding used for reimbursement in that only what was done during a medical encounter should be coded.
- The system didn't have to detect negation. The mention of the absence of a diagnosis or a procedure was also coded.
- The different words that make up an annotation can be very far apart in a clinical note, often without syntactic dependencies. Detecting terms composed of nonadjacent words seemed to be very challenging.
- Spelling mistakes were very rare.

The development and train sets were combined into a single set later referred to as the 'training set'. The training set contained 750 clinical cases with 10,678 and 3,018 annotations of diagnoses and procedures, respectively. The number of annotations made up of nonadjacent words was 14.6% and 39.7% for diagnoses and procedures, respectively.

2.4 Algorithm

We reused the algorithm we developed for the multilingual information extraction task at CLEF eHealth 2018 which consisted of automatically assigning ICD-10 codes to French death certificates [1]. The algorithm was described in details at this occasion and the code is available[1]. The algorithm uses a dictionary-based approach. It takes in input a normalized dictionary and stores it in a tree data structure where each token corresponds to a node. A text to annotate is tokenized after undergoing the same normalization process as the terminology: words are normalized through accents (diacritical marks) and punctuation removal, lowercasing and stopwords removal (if a stopword list is given). The algorithm tries to match a token in a text with a token in the tree using three different techniques: exact match, abbreviation match or a string-distance match based on the Levenshtein distance. The abbreviation match technique uses a dictionary of abbreviations that may be provided in input. When the last token of a term at the leaf of the tree is matched, the algorithm outputs an annotation, meaning a term was found. This algorithm can't detect a term if its words are not in the right order or nonadjacent which occurred frequently in this task.

Dictionaries Interestingly, only 4% of the terms annotated by the medical expert were found in the terminologies after normalization. By comparing the terms from the annotations and the terms from the terminology, a list of the most frequent stopwords was created (no especificado, no especificada...). This list was used to normalize the terminology.

Two dictionaries were constructed:

- The first dictionary (run1) contained only the terms from the annotations of the medical expert from the training set which corresponded to 6,316 terms.
- The second dictionary (run2) was the combination of the first dictionary and the normalized labels of the ICD10-CM terminology. It contained a total of 94,386 terms.

These two dictionaries were tested on the training set to detect and remove terms that could hinder the evaluation metrics. For each term we calculated the number of times it was annotated by the algorithm and by the human annotator. If the ratio between these two numbers was greater than 2, the term was removed from the dictionary. For example, the term "renal" was annotated 67 times by the algorithm but only 2 times by the human expert in the development set. Keeping this term would decrease the precision and the F1 score although the recall would be slightly increased.

[1] https://github.com/scossin/IAMsystem

3 Results

We submitted two runs (one for each dictionary) for the three subtasks. It took less than 5 seconds to annotate the 3,001 documents for the three subtasks on a laptop with Intel Core i7-5700HQ @2.70GH x 8CPUs. We obtained our best F1 score with dictionary 1 for detecting the diagnoses and with dictionary 2 for identifying the procedures. Table 1 shows the performance of our system.

Subtask	Run	MAP	Precision	Recall	F1 score
1	1	0.52	0.82	0.59	0.69
	2	0.51	0.79	0.59	0.68
2	1	0.43	0.66	0.37	0.48
	2	0.49	0.69	0.42	0.52
3	1	-	0.005	0.003	0.005
	2	-	0.006	0.004	0.005
3 (non official)	1	-	0.75	0.52	0.61
	2	-	0.73	0.52	0.61

Table 1. System performance on the CodiEsp test set. MAP: Mean Average Precision.

In subtask 3, an error was detected after publication of the official results. A miscalculation error of the end offset position of each term was fixed and the performance (non official) was reassessed by the organizers.

4 Discussion

The performance was better for the diagnosis subtask than for the procedure one. It was not surprising since the number of terms composed of nonadjacent words were higher in this last task and our algorithm cannot detect such terms. This missing functionality was the main limitation of our system and it probably had a strong impact on our results.

In 2018, the same algorithm obtained a F1-score of 0.786 (precision: 0.794, recall: 0.779) on the task of coding French Death Certificates with the ICD-10 terminology. These better results in 2018 can be explained by a greater number of terms annotated by a medical expert, a shorter text to annotate and no long dependency between words in death certificates.

Adding additional terms (run 2) did not improve the recall in subtask 1 and even reduced the precision. The opposite was observed for procedures (subtask 2) where addition of terms (run 2) improved both recall and precision. The labels in ICD-10 terminologies were of little interest compared to labels from the annotations.

The main advantage of our algorithm is its simplicity and speed. All it needs is a dictionary, a list of abbreviations (optional) and stopwords (optional). The algorithm provides an explanation by outputting start and end position of each detected term. The proposed algorithm can be used as a baseline method for any named entity task and could be integrated into another system to create a more complex approach.

Recently deep learning models based on CNNs and RNNs have shown to achieve better performance on NER tasks in clinical domain [8]. However these models are more data hungry and their training is very costly in terms of computational power. Their advantages are diminished when the number of annotations is low because it is very difficult to predict unseen codes compared to a dictionary-based approach. In this task only 1.7% of the codes were present in the training set.

Further improvement may be possible by using a better curated terminology, a longer list of abbreviations and a phonetic matching strategy to our dictionary based approach.

References

1. Cossin, S., Jouhet, V., Mougin, F., Diallo, G., Thiessard, F.: IAM at CLEF eHealth 2018: Concept Annotation and Coding in French Death Certificates. arXiv:1807.03674 [cs] (Jul 2018), http://arxiv.org/abs/1807.03674, arXiv: 1807.03674
2. Goeuriot, L., Suominen, H., Kelly, L., Miranda-Escalada, A., Krallinger, M., Liu, Z., Pasi, G., Gonzales Saez, G., Viviani, M., Xu, C.: Overview of the CLEF eHealth Evaluation Lab 2020. In: Arampatzis, A. (ed.) Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020), vol. 12260. LNCS (2020)
3. Johnson, S.B., Bakken, S., Dine, D., Hyun, S., Mendonça, E., Morrison, F., Bright, T., Van Vleck, T., Wrenn, J., Stetson, P.: An Electronic Health Record Based on Structured Narrative. Journal of the American Medical Informatics Association : JAMIA **15**(1), 54–64 (2008). https://doi.org/10.1197/jamia.M2131, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2274868/
4. Jovanović, J., Bagheri, E.: Semantic annotation in biomedicine: the current landscape. Journal of Biomedical Semantics **8** (Sep 2017). https://doi.org/10.1186/s13326-017-0153-x, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5610427/
5. Karadeniz, I., Ozgur, A.: Linking entities through an ontology using word embeddings and syntactic re-ranking. BMC Bioinformatics **20** (Mar 2019). https://doi.org/10.1186/s12859-019-2678-8, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6437991/
6. 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 eHealth CLEF 2020. CEUR-WS (2020)
7. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: 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. pp. 102–107. EACL '12, Association for Computational Linguistics, Stroudsburg, PA, USA (2012), http://dl.acm.org/citation.cfm?id=2380921.2380942

8. Wu, Y., Jiang, M., Xu, J., Zhi, D., Xu, H.: Clinical Named Entity Recognition Using Deep Learning Models. AMIA Annual Symposium Proceedings **2017**, 1812–1819 (Apr 2018), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977567/