Classifying clinical case studies with ICD-10 at Codiesp CLEF eHealth 2020 Task 1-Diagnostics

Paula Queipo-Álvarez[1] and Israel González-Carrasco[1]

Computer science Department, Universidad Carlos III de Madrid, Spain
{pqueipo,igcarras}@inf.uc3m.es

Abstract. In this paper, the authors describe the approach and results for the participation in Task 1 (multilingual information extraction) of CLEF eHealth 2020. This work addresses the task of automatically assign ICD-10 codes for Diagnostics clinical case studies in Spanish and English. A dictionary-based approach has been used relying on the terminological resource provided by the organization. The system achieved a mean average precision of 0.115 (precision: 0.866, recall: 0.066).

Keywords: ICD-10 Classification · Clinical case studies · Named-Entity Recognition · Dictionary based.

1 Introduction

In this paper, the authors describe the participation of Human Language and Accessibility Technologies Group (HULAT) at CodiEsp CLEF eHealth 2020, in particular to Task 1: Multilingual Information Extraction. The focus is the ICD-10 coding of Diagnostics, belonging to the sub-task D. CodiEsp is the Clinical Cases Coding in Spanish language Track and is devoted to the automatic coding of clinical cases in Spanish.

1.1 State-of-work

CLEF eHealth has been running these annual evaluation campaigns since 2013 in the Information retrieval, Information Management and Information Extraction. For example, in task of multilingual Information Extraction, named entity recognition (NER), text classification and acronym normalization. In 2018 [1], CLEF organization shared a task to promote automatic clinical coding systems over death reports in the French language (in 2018) and over german non-technical summaries (NPTs) of animal experiments (in 2019).

1.2 Task description

This task utilizes the International Classification of Diseases, 10th revision (ICD-10), which is a terminology resource. The sub-task CodiEsp Diagnosis aims to assign ICD10-CM codes (CIE10 Diagnóstico, in Spanish) to clinical case documents [7]. This terminology is tree-shaped. Annotated codes are minimum 3-character long, and the codes with a greater number of characters are more granular. The organization provided a list of valid codes for this sub-task with their English and Spanish description, and an annotated corpus as well. The evaluation of the automatic coding is against manually generated ICD10 codifications. The motivation of the task is to determine the most competitive approaches for coding this type of documents and generating new clinical coding tools for other languages and data collections.

Task 1: Multilingual Information Extraction, was built upon information extraction tasks from previous years[10] [9]. Information Extraction could be treated as a cascaded named entity recognition with normalization, or a text classification. This task was proposed to explore multilingual approaches, even though the two languages (English and Spanish) were addressed individually.

2 Method

In the following subsections, we describe the corpora, the terminology, the dictionary based approach, and SpaCy Language Processing Pipelines.

The system must predict the codes in both languages, English and Spanish. The dictionary matches the codes with the terms in English and Spanish, and the translation of the Spanish clinical case studies into English was offered. In this case, we recognised entities in each language using the dictionary in both languages.

The automatic classification with ICD-10 codes it is a multi-class and multi-label problem. This can also be considered as a Named-Entitiy Recognition (NER) task.

The code is stored in a Github repositoriy[2].

2.1 Corpora and Terminologies

The CodiEsp corpus is available in several versions. In this work, the third version is considered [6]. The corpus is composed of 1000 clinical case studies annotated by clinical coding professionals meeting strict quality criteria. The clinical case studies have been randomly sampled into three subsets: the train, the development, and the test set. The train set is composed of 500 clinical cases, the development set has of 250 clinical cases each, the test set includes 250 clinical cases, and the background set is composed of 2,751 clinical cases.

Apart from the text files with all the clinical case studies, the annotations for train, development and test sets are provided. The annotation has the following fields: articleID, that refers to the clinical case study, and every ICD10-code found in the clinical case study.

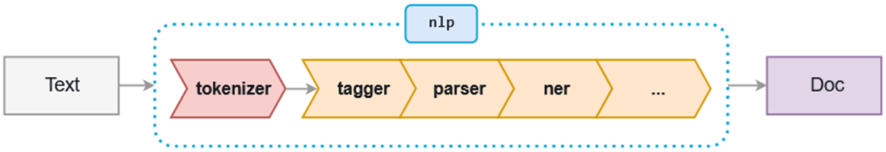

Fig. 1. SpaCy Language Processing Pipeline.

2.2 Dictionary based approach

The organization have provided terminological resources, such as the valid codes for the task[8]. One of the files contains a list of 71486 CIE10-Diagnósticos terms (2018 version) with their description in Spanish and English. Diagnostic codes have the following fields: code, es-description and en-description. To create the dictionary in Python, it was necessary to map the codes to the references and find the frequency of the tags for each code. It is observed that the frequencies differ and the classes are not balanced.

Data preprocessing First of all, it is necessary to join all the clinical case studies in different files for train, development and test to work with them easily. In addition, alphanumeric ICD-10 terms were mapped into different numeric values to avoid errors in the Recognizer. Finally, the clinical case studies were tokenized with SpaCy.

SpaCy Language Processing Pipelines [4] They have been used in several steps. First, to segment text into tokens and produce doc objects that were processed through the pipeline. Secondly, to assign part-of-speech tags using the tagger. It also uses a parser to assign dependency labels. Furthermore, it includes an Entity Recognizer to detect and label named entities.

In the Figure 1 it is shown the structure of the SpaCy Language Processing Pipeline.

In order to use the Entity Recognizer, it is necessary to add Named Entities metadata to doc objects in SpaCy. First, we loaded the English (or Spanish) model. To avoid overlapping of entities, we replaced the default NER module with the dictionary in English (or Spanish). We did this to prevent overlapping of entities.

After that, we detected Named Entities over the test and background set, processing the text and showing its entities.

This procedure has been used with two different models:[5] en_core_web_sm (2.2.5) and es-core-news-sm (2.2.5) in English and Spanish, respectively. These models include convolutional layers, residual connections, layer normalization and maxout non-linearity.

Fuzzy matching With the library fuzzy-wuzzy, it is possible to adjust the maximum (-1) and minimum (50) scores allowed to match between terms. Finally, the predicted terms are saved to evaluate them with the organization script.[3]

2.3 Other approaches

There are other approximations, such as Semantic rules, Transfer learning, RNN-based or BERT-based. Multi-lingual information retrieval (MLIR) is the retrieval of documents in several languages from a query. So it would be interesting to combine English and Spanish in a model.

3 Results & Discussion

In this section, we present the results of one official run. This allows to compare the effectiveness of the classifiers and study the difference in failure analysis.

3.1 Experimental setup

This team submitted predictions for 3001 documents, which includes test files and background files. However, the submission was processed to include only predictions for test files (250 documents) which were considered for computing the metrics.

The evaluation results came from the organization, that used the scripts distributed as part of the Clinical Cases Coding in Spanish language Track (CodiEsp)[3]. All the experiments were performed without cross-validation.

3.2 Test results

The official metric for the subtasks CodiEsp Diagnostic is the mean average precision (MAP). Other metrics computed over a maximum of 1 are: precision, recall and mean average precision for a query of 30 elements (MAP@30).

Table 1. System performance for ICD-10 coding on the test corpus in terms of mean average precision, precision, recall and F-measure. Official evaluation setup.

Parameter	Value
MAP	0.115
MAP30	0.115
Precision	0.866
Recall	0.066
F-measure	0.123

These metrics are computed taking into account only the predictions for the codes presented in the train and development sets. This is because there were codes in the test set that had not been used in the train and validation sets.

In the code search on a set of clinical case studies, precision is the number of correct codes divided by the number of all returned codes, while recall is the number of correct codes divided by the number of codes that should have been returned.

Precision value reaches an 86.6%, whereas in recall only a 6.6%. This means that 86.6% of the codes retrieved were correct, which is impresive due to the difficulty of the taks. Recall measures the fraction of true positives among the predicted results. A low result means a high number of false negatives or codes not detected because of the lexical variability. Our dictionary-based approach was not able to detect the majority of the codes because of a lack of flexibility in the recognition.

Table 2. System performance for ICD-10 coding on the test corpus in terms of mean average precision, precision, recall and F-measure. The test documents without gold labels are ignored for evaluation.

Parameter	Value
MAP codes	0.138
Precision codes	0.935
Recall codes	0.071
F-measure codes	0.132

Also, there are three metrics computed for correct categories. These categories are first three digits of a CIE10-Diagnostic code. For example, codes P96.5 and P96.89 pertains to the category P96.

Table 3. System performance for ICD-10 coding on the test corpus in terms of precision, recall and F-measure. Computed for categories (first 3 digits).

Parameter	Value
Precision categories	0.889
Recall categories	0.074
F-measure categories	0.137

These results are slightly better than the official ones, due to the relaxation of the codes into categories.

We can observe a good result in precision, showing that most of the entities detected where correct, whereas recall results are weak. Also, F-Measure is a good measure when the class distribution is uneven, which is our case. This is an essential problem for this dictionary matching approach. In order to increase the recall, a state-of-the-art approach must be used. Our approach needs to be improved due to the difficulty of the task.

4 Conclusions

This working note presents our contribution to Task 1 of CLEF eHealth competition 2020[7]. The task challenges the automatic assignment of ICD-10 codes for Diagnostics to clinical case studies.

At the previous stages, multi-label classification was tried, but the model did not succeed due to the high number of codes, classes unbalance and a few examples of each class. Then, NER was intended among the clinical case studies to detect Diagnostics, but the model could not predict the code. Due to the lack of results of other approaches, a more traditional dictionary-approach was built. It was able to automatically assign codes to the test set with SpaCy's help.

Evaluation results highlight that our approach was not good enough to detect all the entities presents in the clinical case studies due to the variability of the terms. Lexical variability has impacted recall, that fails to detect terms.

Some improvements would be semantic rules, reduce the number of irrelevant codes, a post-processing filtering phase or including more features. Other upgrades are the treatment of abbreviations and the detection of typos. Also, the future enhancement could be obtained with new terminological and Linguistic resources.

Also, to include deep learning models to infer the categories that have not been seen. This can use and additional corpus to train the model. Finally, it is interesting to implement a system that combines English and Spanish into a multilingual model.

Acknowledgments

This work was supported by the Research Program of the Ministry of Economy and Competitiveness - Government of Spain, (DeepEMR project TIN2017-87548-C2-1-R).

References

1. CodiEsp, https://temu.bsc.es/codiesp/
2. GitHub - pqueipo/Codiesp-CLEF-2020-eHealth-Task1: From Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020.In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum.CEUR Workshop Proceedings (2020), https://github.com/pqueipo/Codiesp-CLEF-2020-eHealth-Task1
3. GitHub - TeMU-BSC/CodiEsp-Evaluation-Script: Evaluation library for CodiEsp Task, https://github.com/TeMU-BSC/CodiEsp-Evaluation-Script
4. Language Processing Pipelines · spaCy Usage Documentation, https://spacy.io/usage/processing-pipelines
5. Models · spaCy Models Documentation, https://spacy.io/models
6. Miranda, A., Gonzalez-Agirre, A., Krallinger, M.: CodiEsp corpus: Spanish clinical cases coded in ICD10 (CIE10) - eHealth CLEF2020 (Apr 2020). https://doi.org/10.5281/zenodo.3758054, https://doi.org/10.5281/zenodo.

3758054, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

7. 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. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2020)
8. Miranda-Escalada, A., Krallinger, M.: CodiEsp codes: list of valid CIE10 codes for the CodiEsp task (Jan 2020). https://doi.org/10.5281/zenodo.3706838, https://doi.org/10.5281/zenodo.3706838, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).
9. Névéol, A., Anderson, R.N., Cohen, K.B., Grouin, C., Lavergne, T., Rey, G., Robert, A., Rondet, C., Zweigenbaum, P.: CLEF eHealth 2017 Multilingual Information Extraction task overview: ICD10 coding of death certificates in English and French. Tech. rep., https://www.cdc.gov/
10. Névéol, A., Cohen, K.B., Grouin, C., Hamon, T., Lavergne, T., Kelly, L., Goeuriot, L., Rey, G., Robert, A., Tannier, X., Zweigenbaum, P.: Clinical Information Extraction at the CLEF eHealth Evaluation lab 2016. Tech. rep., http://quaerofrenchmed.limsi.fr/