Classification of ICD10 Codes with no Resources but Reproducible Code. IMS Unipd at CLEF eHealth Task 1

Giorgio Maria Di Nunzio

Dept. of Information Engineering – University of Padua
giorgiomaria.dinunzio@unipd.it

Abstract. In this paper, we describe the second participation of the Information Management Systems (IMS) group at CLEF eHealth 2018 Task 1. In this task, participants are required to extract causes of death from multilingual death reports (French, Hungarian and Italian) and label them with the correct International Classification Diseases (ICD10) code. We tackled this task by focusing on the reproducible code, that we published last year, which produces a clean dataset that can be used to implement more sophisticated approaches.

1 Introduction

In this paper, we report the experimental results of the second participation of the IMS group to the CLEF eHealth Lab [5], in particular to Task 1: "Multilingual Information Extraction - ICD10 coding" [2]. This task consists in automatically labelling death certificates written in different languages (French, Hungarian, and Italian) with International Classification Diseases (ICD10) codes.

The main goal of our participation to the task this year was to test the effectiveness of the reproducible code made available by [3] which builds a classification system that i) converts raw data into a cleaned dataset following a 'tidyverse' approach[1], ii) implements a set of manual rules to split sentences and translate medical acronyms, and iii) implement a lexicon based classification approach [1].

The contribution of our experiments to this task can be summarized as follows:

- A study of a reproducibility framework to explain each step of the pipeline from raw data to cleaned data;
- An evaluation of the application of a classification system prepared for a language (French) and applied without any additional training or changes to the source code to two different languages (Hungarian and Italian).

We submitted three official runs, one for each language and prepared a number of additional unofficial runs that we will evaluate and compare in order to study the change in performance when adding more information in the pipeline.

[1] https://www.tidyverse.org

Table 1. Expressions in French or punctuation marks used to split a line of a death certificate.

French
avec
sur
par
suite à un[e]
dans un contexte de
après
“,”, “;”, “/”

2 Method

In this section, we summarize the pipeline used in [3] that has been reproduced in this work for each run.

2.1 Pipeline for Data Cleaning

In order to produce a clean dataset, we followed the same pipeline for data ingestion and preparation for all the experiments:

- read a line of a death certificate,
- split the line according to the expression listed in Table 1;
- remove extra white space (leading, trailing, internal);
- transform letters to lower case;
- remove punctuation;
- expand acronyms (if any);
- correct common patterns (if any).

Acronym Expansion Acronym expansion is a crucial step to normalize data and make the death certificate clearer and more coherent with the ICD10 codes. For the French experiments, we used. the original list of 1179 acronyms prepared by a semi-automated approach by [3].

We show the first ten acronym expansions in Table 2. We want to stress the fact that this particular implementation of the expansion selects, in those cases where there is more than once choice (for example “aa”), only the first choice. This is part of our current work in order to improve this step of the pipeline.

2.2 Classification

We used a simple unsupervised lexicon based approach to label each (segment of a) line of a death certificate [1]. The procedure to assign an ICD10 code that does not require any training is the following:

Table 2. Acronym table (fist 10 rows) used to expand acronyms.

acronym	expansion
5-hiaa	acide 5-hydroxyindolactique
5-ht	5-hydroxytryptamine
5-ht	srotonine
a1at	alpha-1-antitrypsine
a1at	a1-antitrypsine
aa	aorte ascendante
aa	affection actuelle
aa	acide amin
aa	antiarthrosique
aaa	anvrisme de l'aorte abdominale

Table 3. Example of classification of a line of a certificate. The definition of the ICD10 labels are shown in Table 4

step	data
line	pneumopathie infectieuse lobaire inférieure droite
terms	pneumopathie, infectieuse, lobaire, inferieure, droite
ICD10 scores	J181 = 7, J13 = 1

- for each term in the (segment of a) line, sum one for each ICD10 label that contains the term,
- for each (segment of a) line compute the score of each ICD10 label;
- group the ICD10 labels that have the maximum score;
- assign the most frequent code within this group.

The score of each label is the sum of the binary weights. In those cases where two or more classes have the same number of entries with the maximum score, the first class in the list is assigned by default. This is another part of the pipeline that requires more effort in order to improve the effectiveness of the classifier. In Table 3, we show an example of the first three steps, while in Table 4 the definition of the ICD10 codes that received the highest score.

3 Experiments and Results

We submitted three official runs, one for each language: French, Hungarian, and Italian. The idea of these experiments was to test the effectiveness of the original French ICD10 classifier on two new languages without any modification to the source code. That is, acronym expansion and sentence splitting are done using French resources. We used only the raw dataset for all the languages.

Table 4. Example of definitions (translitterated) of ICD10 selected in Table 3

ICD10	definition
J13	pneumopathie franche lobaire inferieure
J181	pneumopathie commune lobaire inferieure
J181	pneumopathie infectieuse lobaire aigue
J181	pneumopathie infectieuse lobaire moyenne
J181	pneumopathie infectieuse lobaire superieure
J181	pneumopathie lobaire inferieure
J181	pneumopathie lobaire inferieure aigue
J181	pneumopathie lobaire inferieure bilaterale

Table 5. Results for the official runs

	French			Hungarian			Italian		
	Pre	Rec	F-1	Pre	Rec	F-1	Pre	Rec	F-1
runs	0.6534	0.3963	0.4933	0.7609	0.7482	0.7545	0.5353	0.4844	0.5086
baseline	0.3410	0.2005	0.2525	0.2425	0.1735	0.2023	0.1648	0.1723	0.1685
average	0.7228	0.4102	0.5066	0.8266	0.7830	0.8025	0.8441	0.7606	0.7992
median	0.7981	0.4750	0.5790	0.9221	0.8972	0.9095	0.8995	0.8239	0.8630

3.1 Official Runs

The results of the three experiments are shown in Table 5. The French run performed sufficiently well, and comparable to the results presented in [3]. The F1 measure is close to the average of the results of all the participants in this task. This confirms that a solid clean dataset is a good starting point to build a classifier, even a simple classifier like the one we implemented.

The Hungarian and Italian results are, as we expected, worse than the average scores (much worse for Italian). However, it seems that the Hungarian dataset was in a sense "easier" compared to the our results of our experiments in the Italian subtask. We are going to investigate the reasons for this large difference in performance as future work. Another interesting fact is that, while for the French task Precision was much higher than Recall, for the Hungarian and Italian dataset these two measures seem more "balanced". This may suggest that a better acronym expansion and better sentence splitting may favour Precision over Recall.

3.2 Unofficial Runs

As part of current and future work, we have prepared a set of unofficial runs. A first set of runs study the effect of an alternative weighting scheme, tf-idf instead of binary weighting, another set of runs (for Hungarian and Italian) explore the effectiveness of splitting the sentence with the correct words, see Table 6, as well as expand acronym with the appropriate language. More runs will be created

Table 6. Expressions in Hungarian and Italian or punctuation marks used to split a line of a death certificate.

Hungarian	Italian
a	con
tovább	a causa di
által	per
után a	a seguito di
összefüggésben	conseguentemente a
után	dopo
“,”, “;”, “/”	“,”, “;”, “/”

Table 7. Results for the unofficial runs, tf-idf vs binary weighting

	Hungarian			Italian		
	Pre	Rec	F-1	Pre	Rec	F-1
official	0.7609	0.7482	0.7545	0.5353	0.4844	0.5086
official + tf-idf	0.6870	0.6720	0.6794	0.3642	0.3195	0.3404
binary	0.7559	0.7478	0.7519	0.5353	0.4878	0.5104
binary w/o acronym exp	0.7729	0.7652	0.7690	0.5495	0.5061	0.5269

with additional parameters concerning the multiple label assignment and a better acronym expansion algorithm.

At present time, we have been able to evaluate the effectiveness of some combinations of these parameters. In particular, we tested the binary weighting approach vs the tf-idf approach, using the original French source code (‘inappropriate’ acronyms and sentence splitting), results are shown in the first two lines of Table 7. These results confirms that for Hungarian and Italian the binary weighting approach performs better than tf-idf (the only language that showed some improvement in this task with the tf-idf weights was English [3])

Then, we performed an experiment with binary weights and a ‘correct’ sentence splitting (see Table 6) with or without the French acronym expansion. Results are shown in the last two rows of Table 7. The fact that we used a language specific sentence splitting did not produce any significant change in the performance of the classifier. This is probably due to the fact that the Hungarian and Italian death certificates are much more structured (from a language standpoint) than French ones. For example, we could rarely find complex sentences with words or terms listed in Table 6 in the Italian certificates. It seems that punctuation marks work sufficiently well for these two languages. Moreover, by removing the French acronym expansion, we obtained a slight improvement due to the fact that we removed the noise introduced by a module in the pipeline (the acronym expansion). In this case, results are better in terms of both Precision and Recall compared to the official runs.

4 Final remarks and Future Work

The aim of our second participation to the CLEF eHealth Task 1 was to test the reproducibility of the source code of the lexicon based classifier that was implemented the previous year. The performance of the French run was good and we consider to use it as a baseline to build a new and improved classifier. The application of this classifier to two different language gave interesting results: the results of the Hungarian run was surprisingly high and close to the average of the results of the participant. However, the high value of the median of F1 (close to 90%) suggests that this subtask may be easier than the French one. For the Italian run, we obtained a worse performance the reasons of which we will investigate in a failure analysis.

As current and future work, we are studying

- the adaptation of the pipeline to the two new languages (better split sentence and acronym expansion [4]);
- the possibility to include multiple acronym expansions;
- how to assign multiple labels to the same line (when scores are tied).

References

1. Jacob Eisenstein. Unsupervised learning for lexicon-based classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3188–3194, 2017.
2. A. Névéol, Robert A., F. Grippo, C. Morgand, C. Orsi, L. Pelikán, L. Ramadier, G. Rey, and P. Zweigenbaum. Clef ehealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian. In *CLEF 2018 Evaluation Labs and Workshop: Online Working Notes.* CEUR-WS.org, September 2018.
3. Giorgio Maria Di Nunzio, Federica Beghini, Federica Vezzani, and Geneviève Henrot. A lexicon based approach to classification of ICD10 codes. IMS unipd at CLEF ehealth task 1. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017.
4. Borbála Siklósi and Attila Novák. Detection and expansion of abbreviations in hungarian clinical notes. In Félix Castro, Alexander Gelbukh, and Miguel González, editors, *Advances in Artificial Intelligence and Its Applications*, pages 318–328, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
5. Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, Evangelos Kanoulas, Leif Azzopardi, Rene Spijker, Dan Li, Aurélie Névéol, Lionel Ramadier, Aude Robert, Guido Zuccon, and Joao Palotti, editors. *Overview of the CLEF eHealth Evaluation Lab 2018. CLEF 2018 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science.* Springer, September 2018.