A Lexicon Based Approach to Classification of ICD10 Codes. IMS Unipd at CLEF eHealth Task 1

Giorgio Maria Di Nunzio[1], Federica Beghini[2], Federica Vezzani[2], and Geneviève Henrot[2]

[1] Dept. of Information Engineering – University of Padua
[2] Dept. of Linguistic and Literary Studies – University of Padua
giorgiomaria.dinunzio@unipd.it, fede.beghini92@gmail.com, federicavezzani92@gmail.com, genevieve.henrot@unipd.it

Abstract. In this paper, we describe the participation of the Information Management Systems (IMS) group at CLEF eHealth 2017 Task 1. In this task, participants are required to extract causes of death from death reports (in French and in English) and label them with the correct International Classification Diseases (ICD10) code. We tackled this task by focusing on the replicability and reproducibility of the experiments and, in particular, on building a basic compact system that 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 IMS group that participated for the first time to the CLEF eHealth Lab [8], in particular to Task 1: "Multilingual Information Extraction - ICD10 coding" [11]. This task consists in labelling with International Classification Diseases (ICD10) codes death certificate texts written in English or in French. This work is usually performed by experts in medicine; however, when large volumes of data need to be organized and labelled, manual work is not only expensive but also time consuming and probably not feasible when hundreds of thousands of death certificates need to be classified according to a taxonomy of thousands of codes. For this reason, a possible solution is to approach this task either from a machine learning perspective and/or a natural language processing perspective by using syntactic and/or semantic decision rules [2].

The main goal of our participation to this task was to build a reproducible set of experiments of a system that i) converts raw data into a cleaned dataset, ii) implements a set of manual rules to split sentences and translate medical acronyms, and iii) implement a lexicon based classification approach with the aim of building a sufficiently strong baseline (our initial objective was to achieve a classifier with precision and recall equal 0.5) . We intentionally did not make use of any machine learning approach to improve the accuracy of the classification of death certificates; in fact, the main objective was to build a modular

system that can be easily enhanced in order to make use of the cleaned training data available. For this purpose, we devised a pipeline for processing each death certificate and producing a 'normalized' version of the text. Indeed, death certificates are standardized documents filled by physicians to report the death of a patient but the content of each document contains heterogeneous and noisy data that participants had to deal with [9]. For example, some certificates contain non-diacritized text, or a mix of cases and diacritized text, acronyms and/or abbreviations, and so on.

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

- A reproducibility framework to explain each step of the pipeline from raw data to cleaned data;
- A minimal expert system based on rules to split sentences and translate acronyms;
- Experimenting different weighting approach to retrieve the items in the dictionary most similar to the portion of the certificate of death;
- A simple classification approach to select the ICD code with the highest weight.

For this task, we submitted 2 official English runs plus 3 unofficial English runs and 8 unofficial French runs.

2 Method

In this section, we describe the main aspects of our contribution: the software used to build the reproducibility framework, the data cleaning pipeline, and the classification approach.

2.1 R Markdown for Reproducible Research

The problem of reproducibility in Information Retrieval has been addressed by many researchers in the field in the last years [6, 4, 12]. The main concerns for reproducibility in IR are related to system runs; in fact, even if a researcher uses the same datasets and the same open source software, there are many hidden parameters that make the full reproducibility of the experiment very difficult. For this reason, there are important initiatives in the main IR conferences that support this kind of activity (see for example the open source information retrieval reproducibility challenge at SIGIR[3] or the Reproducibility track at ECIR [5]) as well as in the Natural Language Processing community [1].

During the same time span, the Data Science community has questioned the same issues[4] and has produced interesting solutions from a software point of

[3] `https://github.com/lintool/IR-Reproducibility`
[4] `http://www.nature.com/news/reproducibility-1.17552`

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

English	French
with	avec
due to	sur
that caused	par
sec to	suite à un[e]
on top of	dans un contexte de
also caused by	après
“,”, “;”, “/”	“,”, “;”, “/”

view. The R Markdown framework[5] is now considered one of the possible solutions to document the results of an experiment and, at the same time, reproduce each step of the experiment itself. Following the indications given by [7], we developed the experimental framework in R and publish the source code on github to allow other participants to reproduce our results.[6]

2.2 Pipeline for Data Cleaning

In order to produce a clean dataset, we implemented the following 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 English experiments, we used a manual approach to build the list of expanded acronyms and an automatic approach that gathers acronym from the Web. For the French experiments, we automatically created a list of expanded medical acronyms available on Wikipedia and a manual cleaning of the same list.

Indeed, the automatically creation of a list of acronyms gathered from the Web presents some problems:

- sometimes acronyms have more than one expansion, some of which do not belong to the medical field;

[5] `http://rmarkdown.rstudio.com`
[6] `https://github.com/gmdn/CLEF-eHealth-Task-1`

– some entries contain more than one language, for example English and/or French and/or the Latin expanded acronym;
– some others have some spelling mistakes.

In order to deal with these issues, we referred to the ICD10 dictionary code list which contained a list of diseases and causes of death, to other French dictionaries,[7,8] and to some reliable websites.[9]

Moreover, we removed the wrong definitions and the acronym expansions written in English and in Latin, and we corrected the spelling mistakes concerning some of the accents (especially on the grapheme ¡e¿) and some typos (e.g. "isoniazide" instead of "izoniazide"). Additionally, there were some variants that differed only in the hyphen, e.g. broncho-pulmonaire/bronchopulmonaire, anti-agrégant plaquettaire/anti-agrégant plaquettaire. In these cases, we chose the definition present in the ICD10 dictionary and, if both variants were present, we entered the one that had more occurrences on the Web.

2.3 Classification

We used a simple unsupervised lexicon based approach to label each (segment of a) line of a death certificate [3]. The procedure to assign an ICD10 code that does not require any training is the following:

– for each (segment of a) line compute the score of each entry of the dictionary;
– group the ICD10 codes that have the maximum score;
– assign the most frequent code within this group.

The score of each entry is the sum of the weights of each term either binary weighting (term present or absent) or a term frequency - inverse document frequency (Tf-Idf) approach [10]. 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.

3 Experiments and Results

In our experiments, we implemented:

1. a minimal expert system based on rules to translate acronyms, together with
2. a binary weighting approach or a Tf-Idf approach to retrieve the items in the dictionary most similar to the portion of the certificate of death, and
3. a lexicon based classification approach that selects the most frequent class with the highest weight.

We submitted two official runs for the English raw dataset. Then, we submitted 3 unofficial English runs and 8 unofficial French runs (four for the raw dataset and four for the aligned dataset).

[7] Larousse `http://www.larousse.fr/dictionnaires/francais-monolingue`
[8] Le Trésor de la Langue Française Informatisé `http://atilf.atilf.fr/tlfi.htm`
[9] `http://www.cnci.univ-paris5.fr/medecine/abreviations.html,http://dictionnaire.doctissimo.fr/`

Table 2. Results for the official English runs

EN-ALL	Precision	Recall	F-measure	EN-EXT	Precision	Recall	F-measure
Unipd-run1	0.4963	0.4417	0.4674	Unipd-run1	0.2791	0.0952	0.1420
Unipd-run2	0.3822	0.3405	0.3602	Unipd-run2	0.2917	0.1111	0.1609
average	0.6548	0.5586	0.6017	average	0.3986	0.2749	0.2549
median	0.6459	0.5267	0.5892	median	0.2791	0.2619	0.2740

3.1 Official Runs

For the two official English runs, we pre-processed the raw dataset in the following way:

1. Read the first three fields of the American dictionary (DiagnosisText, Icd1, Icd2, Icd3) and skip lines from 69328 to 69332 since there were some problems with the data format as shown below

   ```
   ...
   LATE EFFECTS TRAUMATIC DUODENAL HEMATOMA;CTS TRAUMATIC ...
   LATE EFFECTS TRAUMATIC DUODENUM HEMORRHAGE;FECTS TRAUMATIC ...
   LATE EFFECTS TRAUMATIC ELBOW HEMATOMA; TRAUMATIC ELBOW HEMORRHAGE; ...
   LATE EFFECTS TRAUMATIC EMPHYSEMATOUS BULLOUS DISEASE;;
   LATE EFFECTS TRAUMATIC EMPHYSEMATOUS LUNG BLEB;
   ...
   ```

2. Index the dictionary using either binary weights or Tf-Idf weights;
3. Build a test run by reading (and cleaning) the causes brutes file and
   - split the sentence according to the following set of patterns: “with”, “due to”, “also due to”, “that caused”, “sec to”, “on top of”,
   - expand each acronym using a table of manually curated acronyms,
4. classify each line by assigning the ICD code with the highest score, if one, or the most frequent code if more than a code matches the line of the death certificate.

The expansion of the acronym was done by manually checking the acronyms in the training data and building a table of expanded acronyms by means of the Web page `https://www.allacronyms.com/_medical`.

The results of the two runs, **Unipd-run1** for the binary weighting approach and **Unipd-run2** for the Tf-Idf weighing approach are reported in Table 2.

The results of the binary weighting run was very close to our expectations, that is to classify correctly almost half of the ICD10 codes (both in terms of Recall and Precision) by just cleaning and normalizing the data without the help of any expert of the field.

The poor result of the Tf-Idf weighting approach on the second run was unexpected. For this reason, we investigated this matter and, thanks to the reproducibility approach, we were able to immediately spot two bugs in the code: 1) we unintentionally selected the Tf weights instead of TfIdf during the

Table 3. Results for the unofficial English runs

EN-ALL	Precision	Recall	F-measure	EN-EXT	Precision	Recall	F-measure
Unipd-run3	0.6104	0.5454	0.5761		0.2167	**0.1032**	**0.1398**
Unipd-run4	0.5015	0.4442	0.4711		**0.2439**	0.0794	0.1198
Unipd-run5	**0.6128**	**0.5474**	**0.5783**		0.1833	0.0873	0.1183
Unipd-run1	0.4963	0.4417	0.4674		0.2791	0.0952	0.1420
Unipd-run2	0.3822	0.3405	0.3602		0.2917	0.1111	0.1609
average	*0.670*	*0.582*	*0.622*	average	*0.405*	*0.267*	*0.267*
median	*0.646*	*0.606*	*0.611*	median	*0.279*	*0.262*	*0.274*

indexing phase, 2) more importantly, we made a mistake in the classification code (step 4 in the above list) that prevented the algorithm to select the most frequent code (it just assigned the first ICD code in the initial list of results). For this reason, we decided to correct the code and submit a second version of Tf-Idf as an unofficial run.

3.2 Unofficial

We also submitted unofficial runs both for French and English with the same original goal but a slightly different approach for the collection of acronyms and the use of transliteration of French diacritics. In particular, we were interested in automatically gathering medical acronyms from a Wikipedia page and manually cleaning the table of expanded acronyms (for example, duplicated entries, both English and French version, wrong diacritics, and so on).

For the expansion of French acronyms, we used the Wikipedia page "Liste d'abréviations en médecine"[10] that contains 1,059 acronyms. After a manual cleaning of the broken/missing/duplicated entries, we produced a table of 1,179 expanded acronyms.

The increase in the number of acronyms is due to the fact that for the same acronym there were several solutions relevant to the medical field. Indeed, we decided to place each variant in a different row with the aim of providing a more complete overview of medical terminology. Furthermore, we applied the same procedure when two acronyms corresponded to the same expansion by keeping both alternatives and positioning them in different rows. Finally, we decided to remove the acronym expansions that were not relevant to the medical field.

For the expansion of the English acronyms, we decided not to use the English Wikipedia list of medical abbreviation page since it is much less informative compared to the French version. Instead, we chose a public Web page that contains 445 common medical abbreviations.[11] For the English unofficial runs, we did not perform any manual corrections of the table of expanded acronyms.

[10] `https://fr.wikipedia.org/wiki/Liste_d\%27abr\'eviations_en_m\'edecine`

[11] `http://www.spinalcord.org/resource-center/askus/index.php?pg=kb.page&id=1413`

Table 4. Results for the unofficial French runs

FR-ALL	Precision	Recall	F-measure	FR-EXT	Precall	Recall	F-measure
Unipd-run6	0.5325	0.3904	0.4505		0.3422	0.2447	0.2854
Unipd-run7	**0.6294**	**0.4684**	**0.5371**		0.3622	0.2505	0.2962
Unipd-run8	0.5326	0.3905	0.4506		0.3839	0.2460	0.2999
Unipd-run9	0.6209	0.4621	0.5299		**0.4052**	0.2503	0.3094
Unipd-run10	0.4383	0.3207	0.3704		0.3197	0.3708	0.3433
Unipd-run11	0.5181	0.3844	0.4413		0.3501	**0.3814**	**0.3651**
Unipd-run12	0.4411	0.3229	0.3728		0.3154	0.3615	0.3369
Unipd-run13	0.5157	0.3827	0.4394		0.3446	0.3702	0.3570
average	*0.4747*	*0.3583*	*0.4059*		*0.3668*	*0.2474*	*0.2921*
median	*0.5411*	*0.4136*	*0.5080*		*0.4431*	*0.2834*	*0.3764*
Unipd-run14	0.4920	0.4203	0.4533		0.3129	0.2577	0.2826
Unipd-run15	0.5941	0.5088	0.5481		0.3323	0.2654	0.2951
Unipd-run16	0.5017	0.4286	0.4623		0.3551	0.2601	0.3003
Unipd-run17	**0.6037**	**0.5170**	**0.5570**		**0.3760**	0.2650	0.3109
Unipd-run18	0.4076	0.3481	0.3755		0.2951	0.3950	0.3378
Unipd-run19	0.4899	0.4193	0.4518		0.3241	**0.4079**	**0.3612**
Unipd-run20	0.4076	0.3480	0.3755		0.2912	0.3922	0.3342
Unipd-run21	0.4884	0.4180	0.4505		0.3198	0.4023	0.3564
average	*0.6479*	*0.5555*	*0.5933*		*0.5051*	*0.3109*	*0.3663*
median	*0.6288*	*0.5396*	*0.5484*		*0.5080*	*0.3330*	*0.4056*

English Run Results A total of three unofficial English runs were submitted:

- **Unipd-run3** a corrected version of the official Tf-Idf run;
- **Unipd-run4** binary weights with automatic acronym expansion;
- **Unipd-run5** Tf-Idf weights with automatic acronym expansion.

The results for the unofficial English runs are reported in Table 3. The first half of the table shows the results of the unofficial runs, while the second half reports the official results for comparison.

French Run Results For the French dataset, we had to lightly change the code that read the aligned and the raw causes since some lines (less than 1% of the data) had some issues with the number of fields (more than expected) and/or contained a semicolon in the death certificate (being the semicolon the separating characters of the fields). See the files available for the reproducibility track for more details.

A total of sixteen unofficial French runs were submitted: eight for the raw dataset, eight for the aligned dataset. For each type of dataset we tried the following settings:

- **Unipd-run6** (raw), **Unipd-run14** (aligned): binary weights, automatic creation of expanded acronyms, without transliteration of diacritics;

– **Unipd-run7** (raw), **Unipd-run15** (aligned): binary weights, automatic creation of expanded acronyms, with transliteration of diacritics;
– **Unipd-run8** (raw), **Unipd-run16** (aligned): binary weights, manually curated expanded acronyms, without transliteration of diacritics;
– **Unipd-run9** (raw), **Unipd-run17** (aligned): binary weights, manually curated expanded acronyms, with transliteration of diacritics;
– **Unipd-run10** (raw), **Unipd-run18** (aligned): Tf-idf weights, automatic creation of expanded acronyms, without transliteration of diacritics;
– **Unipd-run11** (raw), **Unipd-run19** (aligned): Tf-idf weights, automatic creation of expanded acronyms, with transliteration of diacritics;
– **Unipd-run12** (raw), **Unipd-run20** (aligned): Tf-idf weights, manually curated expanded acronyms, without transliteration of diacritics;
– **Unipd-run13** (raw), **Unipd-run21** (aligned): Tf-idf weights, manually curated expanded acronyms, with transliteration of diacritics.

The results for the unofficial French runs are reported in Table 4.

4 Final remarks and Future Work

The aim of our participation was to implement a reproducible lexicon based classifier that can be used as a baseline for further experiments. The performance was sufficiently good and in some cases the classifier achieved a classification performance above 50% both for Recall and Precision which was our initial ideal threshold as a baseline.

Moreover, the preliminary results of the experiments (official and unofficial) have shown interesting differences between the English and French dataset:

– Tf-Idf works better for English while binary weighting performs consistently better for the French dataset;
– For the expansion of the acronym there seems to be a trade-off between manual curation of data and quantity of data gathered from the Web; a lot of noisy data is comparable to a small curated set (see for example Unipd-run3 and Unipd-run5). With lots of data, a round of manual curation allows for small (if not negligible) improvements in terms of accuracy of classification;
– for the French dataset, the normalization of diacritics was a key factor that led to improvements of 10 points percent over the non-normalized version.

Before turning to a more complex system (based on a machine learning approach), we will investigate other forms of data cleaning. In particular, we want to investigate better the problem with diacritics and include an automatic correction of wrong spellings of words (very frequent in the dataset) based, for example, on the Hamming distance among the words of the ICD10 codes.

References

1. Kevin B Cohen, Jingbo Xia, Christophe Roeder, and Lawrence Hunter. Reproducibility in natural language processing: A case study of two r libraries for mining

pubmed/medline. In *In LREC 4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*, pages 6 – 12. European Language Resources Association (ELRA), 2016.

2. Mohamed Dermouche, Vincent Looten, Rémi Flicoteaux, Sylvie Chevret, Julien Velcin, and Namik Taright. ECSTRA-INSERM @ CLEF ehealth2016-task 2: ICD10 code extraction from death certificates. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.*, pages 61–68, 2016.
3. 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.
4. Nicola Ferro. Reproducibility challenges in information retrieval evaluation. *J. Data and Information Quality*, 8(2):8:1–8:4, January 2017.
5. Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors. *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, volume 9626 of *Lecture Notes in Computer Science*. Springer, 2016.
6. Nicola Ferro, Norbert Fuhr, Kalervo Jarvelin, Noriko Kando, Matthias Lippold, and Justin Zobel. Increasing reproducibility in ir: Findings from the dagstuhl seminar on "reproducibility of data-oriented experiments in e-science". *SIGIR Forum*, 50(1):68–82, 2016. http://sigir.org/files/forum/2016J/p068.pdf.
7. Christopher Gandrud. *Reproducible Research with R and R Studio*. Chapman and Hall/CRC, second ed. edition, 2015.
8. Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, João Palotti, and Guido Zuccon, editors. *CLEF 2017 eHealth Evaluation Lab Overview. CLEF 2017 - 8th Conference and Labs of the Evaluation Forum*, Lecture Notes in Computer Science. Springer, 2017.
9. Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, João R. M. Palotti, and Guido Zuccon. Overview of the CLEF ehealth evaluation lab 2016. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, pages 255–266, 2016.
10. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Scoring, term weighting, and the vector space model. In *Introduction to Information Retrieval*, pages 100 – 123. Cambridge, 2008.
11. Aurélie Névéol, Robert N. Anderson, K. Bretonnel Cohen, Cyril Grouin, Thomas Lavergne, Grégoire Rey, Aude Robert, Claire Rondet, and Pierre Zweigenbaum. Clef ehealth 2017 multilingual information extraction task overview: Icd10 coding of death certificates in english and french. In *CLEF 2017 Evaluation Labs and Workshop: Online Working Notes*, CEUR Workshop Proceedings. CEUR-WS.org, 2017.
12. Aurelie Neveol, Kevin Cohen, Cyril Grouin, and Aude Robert. Replicability of research in biomedical natural language processing: a pilot evaluation for a coding task. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 78–84, Auxtin, TX, November 2016. Association for Computational Linguistics.