

A Reproducible Approach with R Markdown to Automatic Classification of Medical Certificates in French

Giorgio Maria Di Nunzio

Dept. of Information Engineering
University of Padua

giorgiomaria.dinunzio@unipd.it

Federica Beghini,

Federica Vezzani, Geneviève Henrot
Dept. of Linguistic and Literary Study

University of Padua

fede.beghini92@gmail.com

federica.vezzani@phd.unipd.it

genevieve.henrot@unipd.it

Abstract

English. In this paper, we report the ongoing developments of our first participation to the Cross-Language Evaluation Forum (CLEF) eHealth Task 1: “Multilingual Information Extraction - ICD10 coding” (Névéol et al., 2017). The task consists in labelling death certificates, in French with international standard codes. In particular, we wanted to accomplish the goal of the ‘Replication track’ of this Task which promotes the sharing of tools and the dissemination of solid, reproducible results.

Italiano. *In questo articolo presentiamo gli sviluppi del lavoro iniziato con la partecipazione al Laboratorio Cross-Language Evaluation Forum (CLEF) eHealth denominato: “Multilingual Information Extraction - ICD10 coding” (Névéol et al., 2017) che ha come obiettivo quello di classificare certificati di morte in lingua francese con dei codici standard internazionali. In particolare, abbiamo come obiettivo quello proposto dalla ‘Replication track’ di questo Task, che promuove la condivisione di strumenti e la diffusione di risultati riproducibili.*

1 Introduction

When researchers use ‘traditional’ methods of scientific publication to describe computational research, we, as readers and researchers, may incur into the so-called ‘reproducible research’ problem (Schwab et al., 2000). For example, a traditional conference paper usually specifies the relevant computations of the main approach, because the limitations of a paper medium prohibit a complete documentation, which would ideally

include experimental data, parameter values, and the source code of the program. Those readers who wish to use the same approach of the paper, hence reproduce the results, must reimplement the whole process, which sometimes may be an unfeasible task. The extreme of reproducibility is ‘replicability’, i.e. a perfect replica of a scientific experiment. The discussion of the difference between replicability and reproducibility is beyond the scope of this paper (Drummond, 2009), and we will just point out that, in general, even in the most accurate replica of an experiment will be done by a different person, in a different lab, using different equipment. Researchers of different areas have identified the necessity for reproducibility, or reproducible research, as an attainable minimum standard for assessing the value of scientific claims (Peng, 2011). As Roger Peng suggests, “one aim of the reproducibility standard is to fill the gap in the scientific evidence-generating process between full replication of a study and no replication. Between these two extreme end points, there is a spectrum of possibilities, and a study may be more or less reproducible than another depending on what data and code are made available”.

Reproducibility matters because the lack of reproducibility in science causes significant issues for science itself, for other researchers in the community, and for public policy. For example, *Nature* published a special issue about “Challenges in Irreproducible Research”¹ where the examined cases showed that there is

[...] a growing alarm about results that cannot be reproduced. Explanations include increased levels of scrutiny, complexity of experiments and statistics, and pressures on researchers. Journals, scientists, institutions and funders all

¹<https://goo.gl/5SxYQJ>

have a part in tackling reproducibility.

Among many other problems, the article showed that most of the drug validation studies (43 out of 67 studies) failed to reproduce. Another important case concerned *Science*, where the Editor-in-Chief retracted in 2015 a study of how canvassers can sway people’s opinions about gay marriage because: “(i) Survey incentives were misrepresented [...], (ii) The statement on sponsorship was false. [...]”² There are also cases of papers retracted by authors themselves because “After carefully re-examining the data presented in the article, they identified that data of two different hospitals got terribly mixed. The published results cannot be reproduced in accordance with scientific and clinical correctness.” as declared in the note of retraction of the paper “Low Dose Lidocaine for Refractory Seizures in Preterm Neonates” (Chakrabarti et al., 2013).

1.1 Reproducible Research in IR and NLP

The problem of reproducibility in Information Retrieval (IR) has been addressed by many researchers in the field in the last years (Ferro et al., 2016b; Ferro, 2017; Neveol et al., 2016). Despite the fact that IR has traditionally been very rigorous about experimental evaluation (the Text REtrieval Conference TREC celebrated the 25th edition in 2016³), many researchers raised some concerns about reproducibility in IR, which are related to system experiments (or runs); in fact, even if a researcher uses the same datasets and the same open source software, there are many parameters and variables hidden in the code that make the full reproducibility of the runs 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 SIGIR⁴ or the Reproducibility track at ECIR (Ferro et al., 2016a)), as well as some Labs at the Cross-Language Evaluation Forum (CLEF) that explicitly have a task on reproducibility, such as CLEF eHealth⁵.

The Natural Language Processing (NLP) community has witnessed the same problem. In 2016, the workshop “Workshop on Research Results Re-

producibility and Resources Citation in Science and Technology of Language” at the Language Resources and Evaluation Conference (LREC) encouraged the discussion and the advancement on the reproducibility of research results and the citation of resources, and its impact on research integrity in the research area of language processing tools and resources. The workshop gathered authors interested in discussing the challenges, the risk factors, the procedures that should be adopted including the new risks raised by the replication articles themselves and their own integrity, in view of the preservation of the reputation of colleagues.

1.2 Contribution

In this paper, we report the current developments of our first participation to the CLEF eHealth Lab (Goeuriot et al., 2017), in particular to Task 1: “Multilingual Information Extraction - ICD10 coding” (Névéol et al., 2017). The task consists in labelling death certificates with standard codes, the International Classification Diseases codes (ICD10). In particular, we wanted to accomplish the goal of the ‘Replication track’ of this task which promotes the sharing of tools and the dissemination of solid, reproducible results (Di Nunzio et al., 2017). Participants of this track had to submit their systems used to produce the experiments, or a remote access to the system, along with instructions on how to install and operate the system. The replication track involved analysts that attempted to replicate a team’s results by running the system supplied on the test data sets, using the team’s instructions.

Therefore, our main objective was to build a modular system that can be easily enhanced in order to make use of the cleaned training data available and to build a reproducible set of experiments of a system that i) converts raw data containing death certificates into a cleaned dataset, ii) implements a set of semi-manual rules to split sentences and translate medical acronyms, and iii) implements a lexicon based classification approach with the aim of building a sufficiently strong baseline (our initial objective was to achieve a classifier performance close to 50%). For this purpose, we devised a pipeline for processing each death certificate and producing a ‘normalized’ version of the text that will be presented in the following sections.

²<https://goo.gl/NWA5gK>

³<http://trec.nist.gov>

⁴<https://goo.gl/CePVzY>

⁵<https://goo.gl/WgkqnZ>

2 R for Reproducible Research

A Tutorial given during the UseR! 2017 conference entitled “Data Carpentry: Open and Reproducible Research with R”⁶ presented an overview of the problems related to (the lack of) reproducible research and the possible solutions in particular when programming with the R Language. In the field of Data Science, the R Markdown framework⁷ is 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 (Gandrud, 2015) and the suggestions discussed by (Cohen et al., 2016), we developed the experimental framework in R and publish the source code on Github⁸ in order to allow other participants to reproduce our results. In particular, in this paper we will focus on the classification of death certificates in French, a part of the work that was partially presented as non-official experiments in the original paper (Di Nunzio et al., 2017).

2.1 Dataset

The CèpiDc corpus was provided by the French institute for health and medical research (INSERM) for the task of ICD10 coding in CLEF eHealth 2017 (Task 1). It consists of free text death certificates collected from physicians and hospitals in France over the period of 2006-2014 (Névél et al., 2017). 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 (Kelly et al., 2016). For example, some certificates contain non-diacritized text, or a mix of cases and diacritized text, acronyms and/or abbreviations, and so on. In Table 1, we show an example of a death certificate of the training set (the English version) split in three lines, Table 1a, and its correct classification with the ICD10 codes, Table 1b. In this case, the last line of the death certificate should be classified with two ICD10 codes (I64 related to acute cerebral issues, and G20 related to Parkinson’s disease). In Table 1c, we show an example of a French death certificate aligned with the cause of death and the ‘standard’ clean text. In both cases, there are issues related with misspellings:

⁶<https://goo.gl/soe9i6>

⁷<http://rmarkdown.rstudio.com>

⁸<https://goo.gl/coCyAe>

the word ‘atrial’ has been written as ‘atril’, as well as many diacritics missing in the French raw text (hemorragie instead of hémorragie).

2.2 Pipeline for Data Cleaning

In order to process the raw death certificate and produce a clean dataset, we implemented the following pipeline for data ingestion: read a line of a death certificate, split the line according to a list of expressions (i.e. “dans un contexte de”, suite à un[e]”, etc.); remove extra white space (leading, trailing, internal); transform letters to lower case; remove diacritics (optional); remove punctuation; expand acronyms (if any); correct common patterns (if any).

The removal of diacritics was surprisingly effective for the French dataset, as discussed in the preliminary experiments (Di Nunzio et al., 2017). For this reason, in this paper we will only show experiments containing this modification. Acronym expansion was also a crucial step to normalize data and make the death certificate clearer and more coherent with the ICD10 codes. For the expansion of French acronyms, we used the Wikipedia page “Liste d’abréviations en médecine”⁹ that contains 1,059 options for acronym expansion. After a manual cleaning of the broken/missing/duplicated entries, we produced a table of 1,179 expanded acronyms.

In this paper, we use a simple semi-automatic step to correct misspellings based on the dictionary of ICD10 codes that was not present in the original experiment. In particular, after cleaning the data and expanding the acronyms, we computed the generalized Levenshtein distance¹⁰ between each token of the death certificate and each token of the dictionary. At the end of this process, we found 4,142 tokens having no match (distance greater than zero) with the ICD10 vocabulary. The terms having more than 10 occurrences in the certificates were hard-coded in the source code, while all the others were automatically substituted on-the-fly.

The vocabulary has 6,295 unique entries, and there are 91,953 lines of 31,682 death certificates to classify.

⁹<https://goo.gl/t41LXn>

¹⁰Given a strings s and t , the Levenshtein distance is the minimal possibly weighted number of insertions, deletions and substitutions needed to transform s into t (so that the transformation exactly matches t).

DocID	YearCoded	LineID	RawText
1	2015	1	PNUEMONIA
1	2015	2	ATRAIL FIBRILLATION
1	2015	6	CVA PARKINSONS DISEASE

(a) Example of death certificate.

DocID	YearCoded	LineID	Rank	ICD10
1	2015	1	1	J189
1	2015	2	1	I48
1	2015	6	1	I64
1	2015	6	2	G20

(b) Example of ICD10 codes for death certificate.

DocID	YearCoded	LineID	RawText	CauseRank	StandardText	ICD10
11	2007	1	hemorragie digestive	1-1	hémorragie digestive	K922
11	2007	2	gastrite	2-1	gastrite	K297
11	2007	5	Pneumopathie , ethylisme chronique , stéatose hépatique	6-1	pneumopathie	J189
11	2007	5	Pneumopathie , ethylisme chronique , stéatose hépatique	6-3	stéatose hépatique	K760
11	2007	5	Pneumopathie , ethylisme chronique , stéatose hépatique	6-2	éthylisme chronique	F102

(c) Example of ICD10 codes for death certificate.

Table 1: Example of death certificate (left) and its correct classification (right) in English Table 1a and 1b. Example of French aligned data in Table 1c.

Table 2: Example of out of vocabulary terms at Levenshtein distance 1.

token	dictionary
alcolique	alcoolique
alcoolo	alcool
arteriopathie	arteriopathie

2.3 Classification rule

The classification of each line of a death certificate uses the approach, proposed by (Eisenstein, 2017), which is performed in the following way: for each line, the score s_i of each entry i of the ICD10 dictionary is computed according to the following sum

$$s_i = \sum_{t_j} w_j \quad (1)$$

which the sum of the weights w_j of each term t_j using binary weighting (one if term present, zero if absent). In those cases where two or more classes have the same score, the first class in the list is assigned by default.

3 Experiments and Results

For the experiments of this paper, we used the ‘raw’ dataset, that is the portion of dataset where a file records the native text entered in the death certificates (referred to as ‘raw causes’ thereafter). System performance was assessed by means of a script provided by the organizers of the Lab; the script computes micro-Precision (the fraction of correct instances among the retrieved instances), micro-Recall (the fraction of relevant instances that have been retrieved over total relevant instances), and micro-F1 measure (the harmonic

mean between micro-Precision and micro-Recall). As requested by the task, these measures were computed for all causes (FR-ALL) in the datasets and for external causes (FR-EXT), where the evaluation is limited to ICD codes addressing a particular type of deaths, called external causes or violent deaths (see the Task overview for more information (Névéol et al., 2017)).

In Table3, we compare the preliminary results of the non-official French experiments submitted in (Di Nunzio et al., 2017) with our ongoing work on cleaning data that makes use of the semi-automatic approach to correct misspellings and different strategies to split the sentences of the death certificate. In particular, we kept the best performing experiment for all causes named **Unipd-run7** which uses binary weights, automatic creation of expanded acronyms and transliteration (removal) of diacritics. The results show the performances on all causes (FR-ALL) as well as the external causes (FR-EXT).

In the new experiment, we tried to vary the approach of splitting the sentences of a death certificate by: non-splitting the sentence (no-split), using only punctuation characters to split like commas, semi-colon, etc. (simplesplit), and using the same strategy of the original experiment (allsplit). We also tried to use the semi-automatic check-spelling (exp) that uses a mix of manual checking for the most common misspelled words (a misspell that occurs more than 10 times in the dataset) and an automatic substitution for all the remaining misspelled words (partialexp).

The experimental results showed that in all cases we could achieve our initial goal that was a classification performance around 0.50 for the F1 measure; moreover, our approach performed bet-

Table 3: Comparison of results with the best performing unofficial French runs and different approaches to certificate segmentation and semi-automatic spell-checking. The average and median performances of all the experiments of the participants of CLEF eHealth Task 1 are reported at the bottom of the table.

	FR-ALL			FR-EXT		
	Precision	Recall	F1	Precision	Recall	F1
Unipd-run7	0.630	0.468	0.537	0.362	0.251	0.296
Unipd-exp-nosplit	0.645	0.400	0.494	0.438	0.220	0.293
Unipd-exp-simplesplit	0.644	0.456	0.534	0.421	0.233	0.300
Unipd-exp-allsplit	0.645	0.483	0.552	0.393	0.253	0.307
Unipd-partialexp-allsplit	0.646	0.484	0.554	0.409	0.255	0.314
average	0.475	0.358	0.406	0.367	0.247	0.292
median	0.541	0.414	0.508	0.443	0.283	0.377

ter than the average and the median score of all the experiments that were submitted to the CLEF eHealth Task 1. This was a bit of a surprise considering that our classification approach does not use any machine learning approach, but it just cleans the data and assigns the most frequent ICD10 code. This is an encouraging result that sets a solid basis of cleaned data on which we can apply more sophisticated NLP techniques, like those used by the best systems like LIMS (see (Zweigenbaum and Lavergne, 2017)) which relied upon dictionary projection and supervised multi-class, single-label text classification using dictionaries and token bigram features (Névéol et al., 2017).

4 Final remarks and Future Work

The aim of this work was to continue the work on the reproducible research approach that can be used as a baseline for further experiments. The performance of the system that uses a semi-manual spell-checking approach improved the baseline set by the original paper. The documentation produced for the reproducibility approach helped us to spot bugs during the implementation phase and we strongly believe that this type of actions should be supported more and more because, as reported by the analysis who tested the systems at CLEF eHealth “[...] still experienced varying degrees of difficulty to install and run the systems. [...] Analysts also report that additional information on system requirements, installation procedure and practical use would be useful for all the systems submitted, although documentation was overall more abundant and detailed compared to last year’s experiments. [...] The results of the experiments suggest that replication is achievable.

However, it continues to be more of a challenge than one would hope.”

References

- Raktima Chakrabarti, Hans-Georg Topf, and Michael Schroth. 2013. Retraction note to: Low dose lidocaine for refractory seizures in preterm neonates. *The Indian Journal of Pediatrics*, 80(6):529–529, Jun.
- Kevin B Cohen, Jingbo Xia, Christophe Roeder, and Lawrence Hunter. 2016. 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).
- Giorgio Maria Di Nunzio, Federica Beghini, Federica Vezzani, and Geneviève Henrot. 2017. A Reproducible Approach with R Markdown to Automatic Classification of Medical Certificates in French. In *CLEF 2017 Evaluation Labs and Workshop: Online Working Notes, Dublin, Ireland, September 11-14, 2017.*, CEUR Workshop Proceedings. 1866.
- C. Drummond. 2009. Replicability is not reproducibility: Nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*.
- Jacob Eisenstein. 2017. 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.
- Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors. 2016a. *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.
- Nicola Ferro, Norbert Fuhr, Kalervo Jarvelin, Noriko Kando, Matthias Lippold, and Justin Zobel. 2016b. Increasing reproducibility in ir: Findings from the dagstuhl seminar on "reproducibility of data-oriented experiments in e-science". *SIGIR Forum*, 50(1):68–82. <http://sigir.org/files/forum/2016J/p068.pdf>.
- Nicola Ferro. 2017. Reproducibility challenges in information retrieval evaluation. *J. Data and Information Quality*, 8(2):8:1–8:4, January.
- Christopher Gandrud. 2015. *Reproducible Research with R and R Studio*. Chapman and Hall/CRC, second ed. edition.
- Lorraine Goeriot, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, João Palotti, and Guido Zuccon, editors. 2017. *CLEF 2017 eHealth Evaluation Lab Overview. CLEF 2017 - 8th Conference and Labs of the Evaluation Forum*, Lecture Notes in Computer Science. Springer.
- Liadh Kelly, Lorraine Goeriot, Hanna Suominen, Aurélie Névéol, João R. M. Palotti, and Guido Zuccon. 2016. 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.
- Aurélie Neveol, Kevin Cohen, Cyril Grouin, and Aude Robert. 2016. 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, Auxtín, TX, November. Association for Computational Linguistics.
- Aurélie Névéol, Robert N. Anderson, K. Bretonnel Cohen, Cyril Grouin, Thomas Lavergne, Grégoire Rey, Aude Robert, Claire Rondet, and Pierre Zweigenbaum. 2017. 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.
- Roger D. Peng. 2011. Reproducible research in computational science. *Science*, 334(6060):1226–1227.
- M. Schwab, N. Karrenbach, and J. Claerbout. 2000. Making scientific computations reproducible. *Computing in Science Engineering*, 2(6):61–67, Nov.
- Pierre Zweigenbaum and Thomas Lavergne. 2017. Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017*.