A Dataset for ICD-10 Coding of Death Certificates: Creation and Usage

Thomas Lavergne
LIMSI, CNRS, Univ. Paris-Sud
Université Paris-Saclay
F-91405 Orsay
lavergne@limsi.fr

Aurélie Névéol
LIMSI, CNRS
Université Paris-Saclay
F-91405 Orsay
neveol@limsi.fr

Aude Robert
CépiDC
80, rue du Général Leclerc
F-94276 le Kremlin-Bicêtre Cedex
aude.robert@inserm.fr

Cyril Grouin
LIMSI, CNRS
Université Paris-Saclay
F-91405 Orsay
grouin@limsi.fr

Grégoire Rey
CépiDC
80, rue du Général Leclerc
F-94276 le Kremlin-Bicêtre Cedex
gregoire.rey@inserm.fr

Pierre Zweigenbaum
LIMSI, CNRS
Université Paris-Saclay
F-91405 Orsay
pz@limsi.fr

Abstract

Very few datasets have been released for the evaluation of diagnosis coding with the International Classification of Diseases, and only one so far in a language other than English. This paper describes a large-scale dataset prepared from French death certificates, and the problems which needed to be solved to turn it into a dataset suitable for the application of machine learning and natural language processing methods of ICD-10 coding. The dataset includes the free-text statements written by medical doctors, the associated meta-data, the human coder-assigned codes for each statement, as well as the statement segments which supported the coder's decision for each code. The dataset comprises 93,694 death certificates totalling 276,103 statements and 377,677 ICD-10 code assignments (3,457 unique codes). It was made available for an international automated coding shared task, which attracted five participating teams. An extended version of the dataset will be used in a new edition of the shared task.

1 Introduction

Over the past decade, biomedical named entity recognition (NER) and concept normalization have been widely covered in NLP challenges. Different types of texts were explored: clinical texts were used in the CMC (Pestian et al., 2007) and the i2b2 NLP Challenges (Uzuner et al., 2007; Uzuner et al., 2011) while the biomedical literature provided material for the BioNLP-Shared Tasks (Kim et al., 2011; Nédellec et al., 2015). Few challenges offered datasets in more than one languages, such as the CLEF ER (Rebholz-Schuhmann et al., 2013) and CLEF eHealth Challenges (Goeuriot et al., 2015)

The assignment of codes from the International Classification of Diseases (ICD) to clinical texts is primarily used for billing purposes but also has a wide range of applications including epidemiological studies (Woodfield et al., 2015), monitoring disease activity (Koopman et al., 2015a), or predicting cancer incidence through retrospective and prospective studies (Bedford et al., 2014). Nevertheless, useful results can only be achieved if ICD code assignment is accurate (Mieno et al., 2016), and studies evidenced that it is a challenging task even when performed manually (Dalianis, 2014).

This is a motivation for creating shareable datasets for ICD coding from natural language text: text corpora annotated with associated ICD codes that can be used to train and evaluate automatic coding systems. Automatic coding has the potential to reduce the cost of physician involvement in the coding process and to increase the consistency of coding.

A potential source of ICD coding datasets comes from death certificates, which are coded in countries around the world according to the World Health Organization (WHO) international standards, using ICD-10. This coding process exists in virtually every country, hence in a large variety of languages. We describe herein the creation of a large-scale ICD coding dataset from death certificates, instantiated in the case of France and the French language. This experience can pave the way for other instantiations in other countries and languages.

Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016), pages 60–69, Osaka, Japan, December 12th 2016.

We first review related work on ICD coding datasets, briefly mentioning associated automated coding methods. We then present the material we started from, the issues we encountered and how we solved them. We describe the resulting data and its use in an international shared task.

2 Related work

In 2007, the Computational Medicine Center (CMC) challenge proposed to identify ICD-9-CM disease codes on a corpus of outpatient chest x-ray and renal procedures (Pestian et al., 2007). In those documents, two sections are identified as more likely to yield codes: 'clinical history' and 'impression'. Both training set and test set are well balanced (respectively 978 and 976 documents). The corpus targeted a subset of only 45 ICD-9-CM codes so that each one of the 94 distinct combination of codes from the test set were seen during the training stage. The best system used a decision tree and achieved a 0.89 F-measure on the test set.

Apart from the CMC challenge, various studies have addressed automatic ICD-10 coding. Koopman et al. (2015a) classified Australian death certificates into 3-digit ICD-10 codes such as *E10* with SVM classifiers based on n-grams and SNOMED CT concepts, and with rules. They also trained SVM classifiers (Koopman et al., 2015b) to find ICD-10 diagnostic codes for death certificates. In contrast to the dataset presented here, they only addressed cancer-related certificates. In addition, they tackled the level of 3-digit ICD-10 codes (e.g., *C00*, *C97*) instead of the full 4-digit level usually required for ICD-10 coding (e.g., *C90.2*). Another important difference is that they focused on the underlying cause of death, i.e., one diagnosis per death certificate. The present dataset keeps all the diagnoses mentioned in each statement of a given death certificate, so that the number of codes to assign to a certificate varies from document to document and is not known *a-priori*. This dataset is intended to support a statement-coding task rather than as a certificate-coding task.

Perotte et al. (2014) took advantage of the presence of ICD-9 codes in the MIMIC-II database along with free text notes. They tested the use of the hierarchical structure of the ICD codes system to improve automatic coding. They compared two coding approaches to assign ICD-9 codes to documents, using SVM classifiers: one took into account the hierarchical structure of ICD-9 codes (hierarchy-based classifier); the other did not (flat classifier). They report higher recall (0.300) and F-measure (0.395) when using the hierarchy-based classifier.

All of this work addressed English language free text. Additionally, ICD-10 coding shared tasks from Japanese clinical records were organized at NTCIR-11 (MedNLP-2) (Aramaki et al., 2014) and NTCIR-12 (MedNLPDoc) (Aramaki et al., 2016). The latter included 200 medical records with an average 7.82 sentences and 3.86 ICD codes per record, totalling 552 distinct codes. However, the inter-annotator agreement was low, with an F-measure of 0.235. The best system obtained an F-measure of 0.348.

We present here the construction and use of a much larger-scale ICD-10 coding dataset in French. Instead of clinical records, it is based on much shorter narratives, viz. death certificates.

3 Material and methods

This section describes the original data; it presents the issues that prevented direct use for a shared task as well as the processing methods we designed to create a dataset suitable for a shared task.

3.1 The coding process at the French WHO collaborative center

Causes of death statistics are essential data to monitor population health, undertake epidemiological studies and international comparisons.

Death certification by a medical practitioner is a mandatory procedure for any death occuring on the French territory. It can be done on a paper certificate or through a secure Web application. In 2007, electronic certification was introduced in France with the objective (among others) to provide a much quicker process for health surveillance and alert systems (Pavillon et al., 2007). Currently, around 12% of death certificates are electronically certified. The system is run on a completely voluntary basis.

Paper death certificates are keyed in by contractors. In this process, contractors may normalize parts of the text to facilitate its subsequent coding; for instance, disease mentions may be replaced with an

equivalent from a standard dictionary.

Causes of death data is centralized at the French Epidemiological Center for the Medical Causes of Death (CépiDc – Inserm). Death certificates are coded with the international software IRIS (Johansson and Pavillon, 2005) in order to assign a code selected from the International Classification of Diseases, tenth revision (ICD-10) to each reported nosologic entity. Then several ICD rules are applied in order to select the so-called underlying cause of death, which is used in most statistics compilations.

Death certificates are now increasingly produced electronically. While this makes the documents more easily available for machine processing, it also creates new challenges. Since electronic certificates are not handled by contractors, their variability of expression is higher than that of transcribed paper certificates; they can also contain spelling errors. Therefore, it is more difficult to handle automatic processing of electronic certificates compared to transcribed certificates which are currently handled by IRIS. This creates an additional motivation for testing state-of-the-art automatic coding methods on modern death certificates, as can be done in a shared task. For these reasons, we used electronic death certificates to create the dataset described herein.

3.2 Data produced by this coding process

In compliance with the World Health Organization (WHO) international standards (Wor, 2011), French death certificates are composed of two parts: Part I is dedicated to the reporting of diseases related to the main train of events leading directly to death, and Part II is dedicated to the reporting of contributory conditions not directly involved in the main death process. According to WHO recommendations, the completion of both parts is free of any automatic assistance that might influence the certifying physician.

In the course of coding practice, the data is stored in different files: a file that records the native text entered in the death certificates (called 'raw causes' thereafter) and a file containing the result of normalizing the text and assigning ICD codes (called 'computed causes' thereafter). An example of 'raw' and 'computed' causes is show below in Table 1.

3.3 Encountered issues

We found that the formatting of the data into raw and computed causes made it difficult to directly relate the codes assigned to original death certificate texts, which would reduce the interest of the data for a shared task. The main issues we identified were:

1. **Outside information needed.** Some coding decisions were made after complementary information was obtained through another channel, such as by contacting the author of the certificate. No trace of this communication is present inside the death certificate itself, hence its contents are not relevant as a source for coding.

2. **Alignment challenge.** The correspondence between the 'computed causes' records in the computed causes file and the statements in the raw causes file could not be easily recovered through the information present in these files. The raw causes file used actual line numbers of the source death certificate (1–4 and 5), but the computed causes file sometimes did not keep the order of the causes as mentioned in the raw causes, and used line numbering that could arbitrarily differ from that of the raw causes. Further more, the text of the computed causes consists of a normalized excerpt of the raw causes text that lead to the specific code assignment. In practice, this means that the specific text strings were related, but often not identical.

The certificates which needed outside information to assign the correct code could be identified through the mention of conditions that prevent a specific code assignment: *décès de cause inconnue* (*unknown cause of death*), *autopsie en cours* (*autopsy requested*) or through the automatic detection of incoherence between the cause mention and the patient age or gender. In those circumstances, a letter is addressed to the doctor, in order to request additional information. Every year, about 1,800 letters are sent and 500 answers received. With this feedback, codes are directly assigned to the corresponding certificates without revising the original text; instead, a free text comment reporting on the supporting correspondance is entered in the coding software. The certificates meeting these criteria were then removed from the dataset.

The alignment issue required to find a method to *align* the source statements with the computed causes records. We describe this method in the next section.

3.4 Pre-processing of death certificate through alignment

The goal of the alignment process is to obtain (statement, code) pairs, where the statement includes the original text and its associated meta-information, as per the raw causes file, and the code is one of those which should be assigned to this statement as per the computed causes file, together with the associated normalized text. Input statements with multiple codes are repeated in multiple (statement, code) pairs.

A sample document is presented in table 1. This example illustrates different types of difficulty of the alignment step:

- cause order is reversed (e.g., *choc septique* appears in line 1 of the raw causes but in line 3 of the computed causes),
- multiple causes are merged on a single raw line (e.g. *peritonite stercorale* and *perforation colique* on line 2),
- different capitalization and stopwords (e.g. see line 3 of aligned causes),
- different spelling. There is no occurrence of this in our sample document; however, a raw cause such as *bactériémie à K. pneumoniae* would be normalized to *bactériémie klebsiella pneumoniae*, using a variant of the name of the bacteria involved in the reported infection.

Table 1: A sample document from the CépiDC French Death Certificates Corpus: alignment of the raw causes and computed causes. English translations for each text line are provided in footnotes.

data type	line	text	normalized text	ICD codes
Raw causes	1	choc septique[1]		-
	2	peritonite stercorale sur perforation colique[2]		-
	3	Syndrome de détresse respiratoire aiguë[3]		-
	4	defaillance multivicerale[4]		-
	5	HTA[5]		-
Computed causes	1		defaillance multivicerale	R57.9
	2		syndrome détresse respiratoire aiguë	J80.0
	3		choc septique	A41.9
	4		peritonite stercorale	K65.9
	5		perforation colique	K63.1
	6		hta	I10.0
Aligned causes	1	choc septique	choc septique	A41.9
	2	peritonite stercorale sur perforation colique	peritonite stercorale	K65.9
	2	peritonite stercorale sur perforation colique	perforation colique	K63.1
	3	Syndrome de détresse respiratoire aiguë	syndrome détresse respiratoire aiguë	J80.0
	4	defaillance multivicerale	défaillance multiviscérale	R57.9
	5	HTA	hta	I10.0

Our alignment method relied both on the order that causes and codes occurred in the files and on string similarity between the texts of raw and computed causes. More specifically, the principles we followed to reconcile raw and computed causes were the following:

[1] *septic shock*
[2] *colon perforation leading to stercoral peritonitis*
[3] *Acute Respiratory Distress Syndrome*
[4] *multiple organ failure*
[5] *HBP: High Blood Pressure*

- Alignments have the form $(0, 1) \rightarrow m$
- All computed causes must be supported by an input statement
- No alignment should have the form $n \rightarrow m$. However, some death certificates contain separate input statements which must be taken as a whole to produce a relevant code. An example is the set of two lines *1. Strangulation au lien* (*ligature strangulation*) and *2. Suicide* which is coded as X70.9 (*Intentional self-harm by hanging strangulation and suffocation home during unspecified activity*), where the suicide must be coded by taking into account the specific circumstance that lead to it (here, strangulation). In such cases we kept the input statements separate. The most generic statement (e.g. *suicide*) was considered inconclusive and did not receive a code assignment while the 'head' statement (e.g. *ligature strangulation*, which provided the defining information for code assignment) was aligned with the output code.

To align the statements, we used a model originally intended for bilingual word alignment in parallel sentences: a log-linear reparameterization of the IBM2 model (Dyer et al., 2013). The alignments were produced from the computed clauses without allowing for null alignment in order to satisfy our constraints, and with a Dirichlet prior to favor diagonal alignments.

The model underperforms on multi-word segments as it relies on co-occurrence counts of raw and computed causes, which are very sparse. To overcome this problem, both causes were pre-processed by removing stopwords and applying stemming. Next, the Damerau-Levenshtein distance between two segments was linearly combined with the occurrence count to act as a prior on the alignment probabilities.

4 Results

We applied the above-described methods to the 2006–2013 death certificates created by the electronic work-flow and describe the resulting data and its usage.

4.1 Corpus characteristics

Table 2 presents the fields found in each line of the produced dataset. One line is produced for each (input line, output code) pair. Some input lines have no associated output code: the corresponding values are empty. As explained in Section 3.4, this also occurs when two "raw cause" input lines need to be considered together to be coded. In that case, only one of them has an associated code.

The dataset was split into training and test sets: the training set contains statements of years 2006–2012, and the test set contains statements of year 2013. We now provide more detail on the training set.

The distribution of statement length in tokens, after stop-word removal (French stop words of the NLTK toolkit), is shown on Figure 1a. It shows that statement length follows a Zipfian distribution from length 2 to length 31. Statements over 20 tokens are rare (455 = 0.17%), over 10 tokens too (9538 = 3.6%). The maximum length of a statement is 120 tokens.

Figure 1b shows the most frequent codes. The top five are R092 (Respiratory arrest), A419 (Septicaemia, unspecified), R688 (Other specified general symptoms and signs), I10 (Essential (primary) hypertension), I509 (Heart failure, unspecified). These top diagnoses, as well as those in the rest of the figure, display a mixture of very general diagnoses (*unspecified*, *other*) and most frequent causes of death (infection, hypertension, pneumonia, cancer, etc.).

ICD-10 is divided into 21 chapters. Figure 1c shows the number of codes in each chapter in the training set. The most represented chapters are Chapters IX (codes I00–I999, Diseases of the circulatory system), II (C00–D489, Neoplasms), XVIII (R00–R999, Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified), etc. Figure 1d shows the number of occurrences of each age group for each chapter. A few chapters have a skewed distribution of age groups: P00-P969 (Certain conditions originating in the perinatal period: 99.8% for age group 0), Q00-Q999 (Congenital malformations, deformations and chromosomal abnormalities, 55.4% for age group 0), and O00-O999 (Pregnancy, childbirth and the puerperium: 15.4% for age group 25, 46.3% for 30, 34.6% for 35).

Field	Contents
DocID	Death certificate ID.
YearCoded	Year the death certificate was processed by the human coder.
Gender	gender of the deceased
Age	Age at the time of death, rounded to the nearest five-year age group.
LocationOfDeath	Location of death, according to the following categories: 1 = Home; 2 = Hospital; 3 = Private Clinic; 4 = Hopice, Retirement home; 5 = Public place; 6 = Other Location.
LineID	Line number within the death certificate. Note that if a statement is assigned multiple ICD10 codes, it is repeated for each code, each time with the same LineID.
RawText	Raw text entered in the death certificate.
IntType	Type of time interval the patient had been suffering from coded cause, according to the following categories: 1 = minutes; 2 = hours; 3 = days; 4 = months; 5 = years.
IntValue	Length of time the patient had been suffering from coded cause; for example, if the patient had been experiencing the cause for 6 months, *IntValue* should be 6 and *IntType* should be 4.
CauseRank	Rank of the ICD10 code assigned by the human coder. The rank (e.g., *2-1*) is composed of two items found in the original CausesCalculees file: the number of the line (NumLigne, e.g., *2*) followed by the rank of the cause in that line (RangCause, e.g., *1*).
StandardText	Dictionary entry or excerpt of the raw text that supports the selection of an ICD10 code.
ICD10	Gold standard ICD10 code.

Table 2: Fields in each row of the dataset. The last three fields are the output of the coding process.

4.2 Use in a shared task

The resulting dataset was used in an international shared task (Névéol et al., 2016). The certificates corresponding to year 2006-2012 were used as a training set (N=65,844) while certificates corresponding to the year 2013 were used as a test set (N=27,850). A small number of codes (N=244, about 10% of the unique codes in the test set) in the test set were unseen in the training set. Five teams from three countries submitted a total of seven runs for this task. Participating teams used methods relying either on knowledge-base linking or statistical machine learning. Table 3 shows the performance of the official runs, compared to a baseline run, which consisted in assigning codes to lines in the test set when an exactly identical line was also found in the training set. When the line occurred multiple times in the training set, the most frequent code was selected. It can be seen from the table that all runs submitted by participants outperformed the baseline by at least 20 points in F-measure, thus demonstrating that the state of the art in ICD10 coding is quite advanced.

We examined the relative difficulty of finding each expected statement code for the submitted systems: for each death certificate statement and expected code for this statement, we counted the number of systems which correctly found this code. Figure 2 shows the results.

We found out that among the 110767 distinct entries of the test dataset, 29100 were easy to find: all systems found the correct answer; 25215 were fairly easy: all but one system found them; 20743 were less easy (3 systems); 15933 were harder (2 systems); 10685 were rather hard: only one system found them; and 7714 were hard: no system found them at all. The latter may help identify to difficult, hence interesting problems, such as codes which need to refer to the broader context of the full death certificate, beyond the current individual statement, to be assigned properly. They may also point at cases where human coding might not be correct.

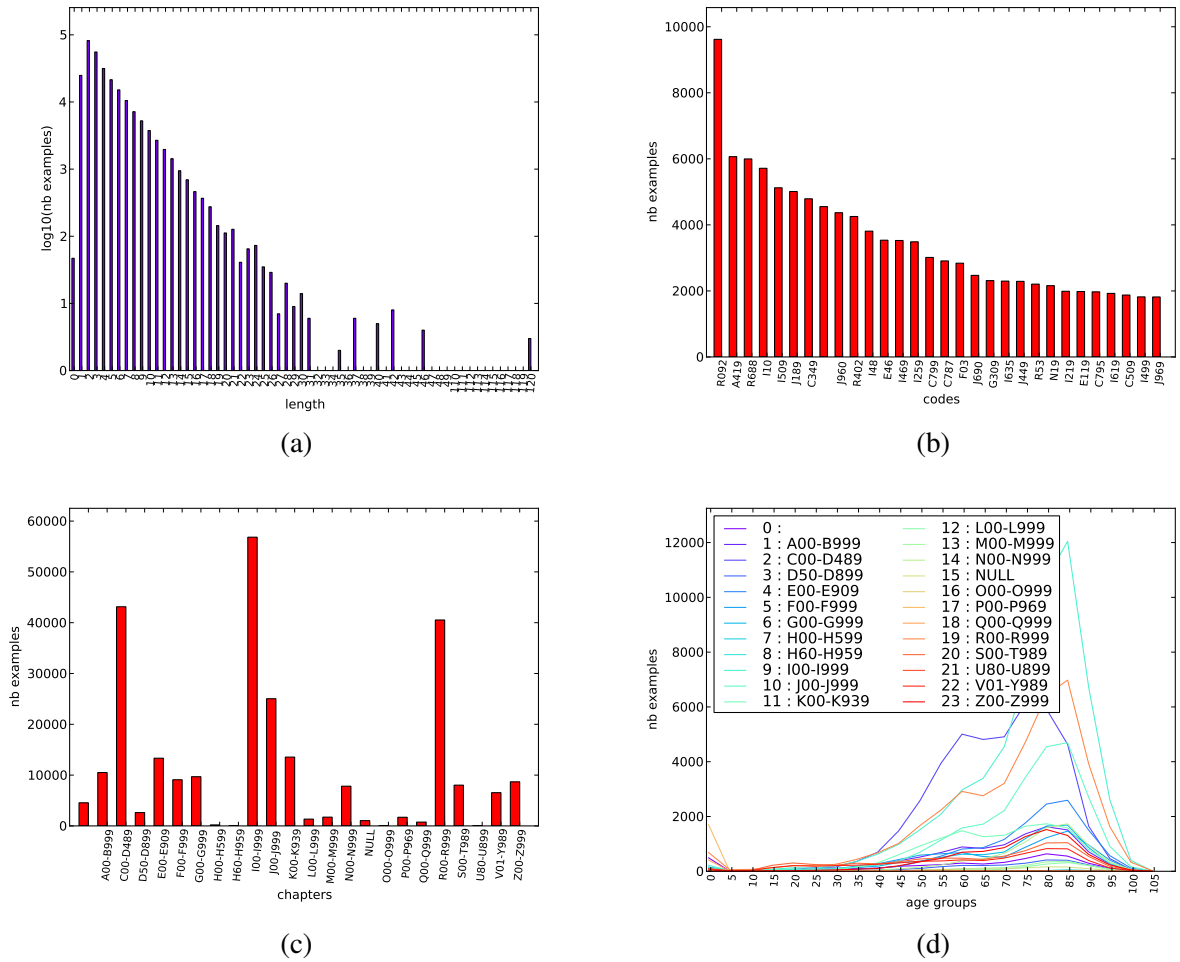

Figure 1: Statistics on training set. (a) Distribution of statement length (after normalization, log-y scale). (b) Most frequent codes. (c) Distribution of ICD-10 chapters. (d) Occurrences of age groups for each chapter.

5 Discussion

With 93,694 death certificates totalling 276,103 statements and 377,677 ICD-10 code assignments (3,457 distinct codes), the size of the presented dataset is comparable to the largest so far on English (Perotte et al., 2014) (22,815 discharge summaries and 215,826 ICD9 codes (5,030 distinct codes)), and is several orders of magnitude above the other ICD coding datasets we identified (Pestian et al., 2007; Aramaki et al., 2014; Aramaki et al., 2016).

An important difference though is that the present dataset consists of death certificate statements, whereas the other cited datasets are made of clinical records such as discharge summaries. Death certificate statements are fairly short and focused on nosologic entities, whereas clinical records are usually longer and mention a broader set of entities and events. Medical records exhibit a large range of sizes however: for instance, texts in the MedNLPDoc dataset (Aramaki et al., 2016) contained on average 7.82 sentences.

A consequence of the difference between death certificate statements and for instance discharge summaries is that death certificate statement coding might be more easily addressed as a text classification task, whereas clinical record coding may need to rely on a step of entity detection and normalization methods to identify more relevant pieces of information before ICD coding proper. This makes the clinical record coding task more difficult and explains the lower F-measures obtained in that context.

Future plans include the extension of the present dataset with death certificates of more recent years

Table 3: System performance for ICD10 coding on the death certificate test corpus. A * symbol indicates statistically significant difference of a run with the runs ranked before and after it, according to a Student test.

Team	TP	FP	FN	Precision	Recall	F-measure
TeamA-run2*	88497	11423	20321	0.886	**0.813**	**0.848**
TeamA-run1*	87404	10823	21414	**0.890**	0.803	0.844
TeamB-run2*	71319	9479	37499	0.882	0.655	0.752
TeamB-run1*	66954	15605	41864	0.811	0.615	0.700
TeamC-run1*	72192	31480	36626	0.696	0.663	0.680
TeamD-run1*	61874	19002	46984	0.765	0.569	0.652
TeamE-run1*	57256	40650	51562	0.585	0.526	0.554
Baseline-Zipf-Top1*	26688	23610	82130	0.531	0.245	0.336

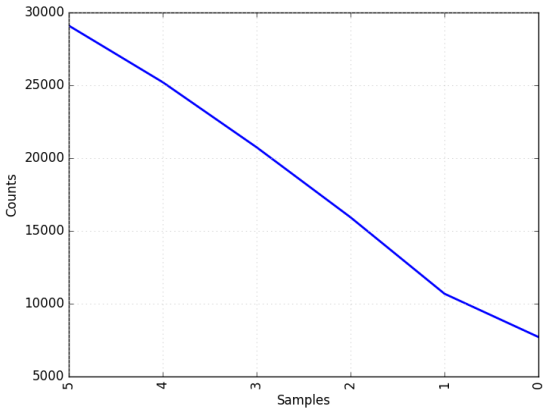

Figure 2: Distribution of coding difficulty based on system results, based on the best run of each of the five participating teams. Samples = number of systems which found the expected code for a statement. Counts = number of (statement, code) pairs found by a given number of systems.

as they are processed by human coders. In an upcoming edition of the shared task, he dataset described herein will be used as a training set while more recent data will be offered as a test set. This time-ordered distribution of certificates in the datasets is guided by the practical use case of coding death certificates, where historical data is available to coders who then need to work with current data. The goal of this series of shared tasks is to engage the community in the development of ICD-10 coding methods that can then be integrated to coders work flow as coding assistance and productivity enhancing tools.

We also plan to include additional languages in future datasets, as other WHO collaborating centers express their interest in this enterprise. We hope that the development of a multilingual ICD-10 coding dataset will foster the development of portable methods that can be easily adapted to several languages.

6 Conclusion

This paper presents a new dataset for ICD-10 coding based on death certificates in French. This is a large dataset comprising death certificate statements in a language other than English as well as rich metadata and professionally assigned gold-standard ICD10 codes. The preparation of the dataset involved the use of complex alignment techniques to ensure the quality of the text-code pairings. It was shown to be a suitable tool for evaluating the state of the art in ICD-10 coding in an international shared task. In future work we plan to enhance the dataset with newer data for French as well as other languages in order to foster global approaches to ICD10 coding.

7 Acknowledgements

This work was partially funded by the European Union's Horizon 2020 Marie Sklodowska Curie Innovative Training Networks—European Joint doctorate (ITN-EJD) under grant agreement No:676207, Methods in Research on Research (MiRoR).

This work was supported in part by the French National Agency for Research under the CABeRneT[6] ANR-13-JS02-0009-01 grant.

References

Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. 2014. Overview of the NTCIR-11 MedNLP-2 task. In *Proceedings of the 11th NTCIR Conference*, Tokyo Japan.

Eiji Aramaki, Mizuki Morita, Yoshinobu Kano, and Tomoko Ohkuma. 2016. Overview of the NTCIR-12 MedNLPDoc task. In *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*, Tokyo Japan.

Ronald L. Bedford, Spencer G. Lourens, Charles F. Lynch, Brian J. Smith, and R. William Field. 2014. Utility of death certificate data in predicting cancer incidence. *Am J Ind Med*, 57(2):153–62.

Hercules Dalianis. 2014. Clinical text retrieval - an overview of basic building blocks and applications. In Georgios Paltoglou, Fernando Loizides, and Preben Hansen, editors, *Professional Search in the Modern World: COST Action IC1002 on Multilingual and Multifaceted Interactive Information Access*, pages 147–165. Springer International Publishing, Cham.

Chris Dyer, Victor Chahuneau, and A. Noah Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648. Association for Computational Linguistics.

Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Leif Hanlen, Aurélie Névéol, Cyril Grouin, João Palotti, and Guido Zuccon, 2015. *Overview of the CLEF eHealth Evaluation Lab 2015*, pages 429–443. Springer International Publishing, Cham.

Lars Age Johansson and Gérard Pavillon. 2005. IRIS: A language-independent coding system based on the NCHS system MMDS. In *WHO-FIC Network Meeting*, Tokyo, Japan.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011. Overview of BioNLP shared task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1–6, Portland, Oregon, USA, June. Association for Computational Linguistics.

Bevan Koopman, Sarvnaz Karimi, Anthony Nguyen, Rhydwyn McGuire, David Muscatello, Madonna Kemp, Donna Truran, Ming Zhang, and Sarah Thackway. 2015a. Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC Med Inform Decis Mak*, 15:53.

Bevan Koopman, Guido Zuccon, Anthony Nguyen, Anton Bergheim, and Narelle Grayson. 2015b. Automatic ICD-10 classification of cancers from free-text death certificates. *Int J Med Inform*, 84(11):956–965, November.

Makiko Naka Mieno, Noriko Tanaka, Tomio Arai, Takuya Kawahara, Aya Kuchiba, Shizukiyo Ishikawa, and Motoji Sawabe. 2016. Accuracy of death certificates and assessment of factors for misclassification of underlying cause of death. *J Epidemiol*, 26(4):191–8.

Claire Nédellec, Jin-Dong Kim, Sampo Pyysalo, Sophia Ananiadou, and Pierre Zweigenbaum. 2015. BioNLP Shared Task 2013: Part 1. *BMC Bioinformatics*, 16(Suppl 10), July.

A Névéol, KB Cohen, C Grouin, T Hamon, T Lavergne, L Kelly, L Goeuriot, G Rey, A Robert, X Tannier, and P Zweigenbaum. 2016. Clinical information extraction at the CLEF eHealth evaluation lab 2016. In *CLEF 2016 Online Working Notes*, pages 28–42. CEUR-WS.

Gérard Pavillon, Patrick Coilland, and Éric Jougla. 2007. Mise en place de la certification électronique des causes médicales de décès en France : premier bilan et perspectives [Implementation of the electronic certification of medical causes of death in France: first results and propects]. *Bulletin épidémiologique hebdomadaire*, 35-36:306–308, Sep 18.

[6]CABeRneT: Compréhension Automatique de Textes Biomédicaux pour la Recherche Translationnelle

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *J Am Med Inform Assoc*, 21(2):231–7.

John P. Pestian, Chris Brew, Pawel Matykiewicz, DJ Hovermale, Neil Johnson, K. Bretonnel Cohen, and Wlodzislaw Duch. 2007. A shared task involving multi-label classification of clinical free text. In *Biological, translational, and clinical language processing*, pages 97–104, Prague, Czech Republic, June. Association for Computational Linguistics.

Dietrich Rebholz-Schuhmann, Simon Clematide, Fabio Rinaldi, Senay Kafkas, ErikM. van Mulligen, Chinh Bui, Johannes Hellrich, Ian Lewin, David Milward, Michael Poprat, Antonio Jimeno-Yepes, Udo Hahn, and JanA. Kors. 2013. Entity recognition in parallel multi-lingual biomedical corpora: The CLEF-ER laboratory overview. In Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, and Benno Stein, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138 of *Lecture Notes in Computer Science*, pages 353–367. Springer Berlin Heidelberg.

Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*, 14:550–563.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc*, 18(5):552–556, Sep-Oct. Epub 2011 Jun 16.

Rebecca Woodfield, Ian Grant, UK Biobank Stroke Outcomes Group, UK Biobank Follow-Up and Outcomes Working Group, and Cathie L. M. Sudlow. 2015. Accuracy of electronic health record data for identifying stroke cases in large-scale epidemiological studies: A systematic review from the UK biobank stroke outcomes group. *PLoS One*, 10(10):e0140533.

World Health Organization, 2011. *ICD-10. International Statistical Classification of Diseases and Related Health Problems. 10th Revision. Volume 2. Instruction manual*.