CEUR-WS.org/Vol-2125/invited_paper_18.pdf

CLEF eHealth 2018 Multilingual Information Extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italian

Aurélie Névéol[1], Aude Robert[2], Francesco Grippo[3], Claire Morgand[2], Chiara Orsi[3], László Pelikán[4], Lionel Ramadier[1], Grégoire Rey[2], and Pierre Zweigenbaum[1]

[1] LIMSI, CNRS, Université Paris-Saclay, Orsay, France
firstname.lastname@limsi.fr
[2] INSERM-CépiDc, Le Kremlin-Bicêtre, France
firstname.lastname@inserm.fr
[3] ISTAT, Italy
frgrippo@istat.it and chiara.orsi@istat.it
[4] KSH, Hungary
laszlo.pelikan@ksh.hu

Abstract. This paper reports on Task 1 of the 2018 CLEF eHealth evaluation lab which extended the previous information extraction tasks of ShARe/CLEF eHealth evaluation labs. The task continued with coding of death certificates, as introduced in CLEF eHealth 2016. This large-scale classification task consisted of extracting causes of death as coded in the International Classification of Diseases, tenth revision (ICD10). The languages offered for the task this year were French, Hungarian and Italian. Participant systems were evaluated against a blind reference standard of 11,932 death certificates in the French dataset 21,176 certificates in the Hungarian dataset and 3,618 certificates in the Italian dataset using Precision, Recall and F-measure. In total, fourteen teams participated: 14 teams submitted runs for the French dataset, 5 submitted runs for the Hungarian dataset and 6 for the Italian dataset. For death certificate coding, the highest performance was 0.838 F-measure for French, 0.9627 for Hungarian and 0.9524 for Italian.

Keywords: Natural Language Processing; Entity Linking, Text Classification, French, Biomedical Text

1 Introduction

This paper describes an investigation of information extraction and normalization (also called "entity linking") from French, Hungarian and Italian-language health documents conducted as part of the CLEF eHealth 2018 lab [1]. The task addressed is the automatic coding of death certificates using the International Classification of Diseases, 10th revision (ICD10) [2]. This is an essential task in epidemiology. The determination of causes of death directly results in the production of national death statistics. In turn, the analysis of causes of death at a global level informs public health policies.

In continuity with previous years, the methodology applied is the shared task model[3].

Over the past five years, CLEF eHealth offered challenges addressing several aspects of clinical information extraction (IE) including named entity recognition, normalization [4–7] and attribute extraction [8]. Initially, the focus was on a widely studied type of corpus, namely written English clinical text [4, 8]. Starting in 2015, the lab's IE challenge evolved to address lesser studied corpora, including biomedical texts in a language other than English i.e., French [5]. This year, we continue to offer a shared task based on a large set of gold standard annotated corpora in French with a coding task that required normalized entity extraction at the sentence level. We also provided an equivalent dataset in Hungarian, and a synthetic dataset for the same task in Italian.

The significance of this work comes from the observation that challenges and shared tasks have had a significant role in advancing Natural Language Processing (NLP) research in the clinical and biomedical domains [9, 10], especially for the extraction of named entities of clinical interest and entity normalization.

One of the goals for this shared task is to foster research addressing multiple languages for the same task in order to encourage the development of multilingual and language adaption methods.

This year's lab suggests that the task of coding can be addressed reproducibly with comparable performance in several European languages without relying on translation. Furthermore, a global method addressing three languages at once opened interesting perspective for multi-lingual clinical NLP [11].

2 Material and Methods

In the CLEF eHealth 2018 Evaluation Lab Task 1, three datasets were used. The French dataset was supplied by the CépiDc[1], the Hungarian dataset was supplied by KSH[2] and the Italian dataset was supplied by ISTAT[3]. All three datasets refer to the International Classification of Diseases, tenth revision (ICD10),a reference classification of about 14,000 diseases and related concepts managed by the World Health Organization and used worldwide, to register causes of death and reasons for hospital admissions. Further details on the datasets, tasks and evaluation metrics are given below.

2.1 Datasets

The CépiDc corpus was provided by the French institute for health and medical research (INSERM) for the task of ICD10 coding in CLEF eHealth

[1] Centre d'épidémiologie sur les causes médicales de décès, Unité Inserm US10, `http://www.cepidc.inserm.fr/`.

[2] Központi Statisztikai Hivatal, `https//www.ksh.hu/`.

[3] Istituto nazionale di statistica, `http://www.istat.it/`.

2018 (Task 1). It consists of free text death certificates collected electronically from physicians and hospitals in France over the period of 2006–2015 [12].

The KSH-HU corpus was provided by the Hungarian central statistical office (KSH). It consists of a sample of randomly extracted free text death certificates collected from doctors in Hungary for the year of death 2016. There is no electronic certification in this country, so in contrast to the French corpus, this corpus contains only deaths reported using paper forms (and then transcribed electronically).

The ISTAT-IT corpus was provided by the Italian national institute of statistics (ISTAT). To better preserve confidentiality, the corpus was fabricated based on real data. Indeed, the fake certificates were created from authentic death certificates corresponding to different years of coding. The lines of a synthetic document each came from a different certificate, while ensuring topical coherance and preserving the chain of causes of death (line 1 of a synthetic certificate was created using line 1 of a real certificate). The coherence of age, sex and causes referred were also preserved. The synthetic certificates were then coded as if they reported a real death for 2016. To summarize, this synthetic corpus provides a realistic simulation of language and terminology found in Italian death certificates, together with official coding. Up to 90 percent of the corpus contains terminology completely recognized by the Italian dictionary but it also offers examples of language that cannot be automatically recognized by the Italian system : linguistics variants, new expressions and spelling mistakes in the text for instance. A characteristic of the Italian dictionary is the poverty of labels associated with the ICD-10 codes for external causes (including certificates reporting surgery), which must be reviewed manually by the coding team.

Dataset excerpts. Death certificates are standardized documents filled by physicians to report the death of a patient. The content of the medical information reported in a death certificate and subsequent coding for public health statistics follows complex rules described in a document that was supplied to participants [12]. Tables 1, 2 and 3 present excerpts of the corpora that illustrate the heterogeneity of the data that participants had to deal with. While some of the text lines were short and contained a term that could be directly linked to a single ICD10 code (e.g., “choc septique”), other lines could contain non-diacritized text (e.g., “peritonite...” missing the diacritic on the first “e”), abbreviations (e.g., “BPCO” instead of “broncopneumopatia cronica ostruttiva”). Other challenges included run-on narratives or mixed text alternating between upper case non-diacritized text and lower-case diacritized text.

Descriptive statistics. Table 4 present statistics for the specific data sets provided to participants. For two of the languages, the dataset construction was time-oriented in order to reflect the practical use case of coding death certificates,

Table 1. A sample document from the CépiDC French Death Certificates Corpus: the raw causes (Raw) and computed causes (Computed) are aligned into line-level mappings to ICD codes (Aligned). English translations for each *raw* line follow: 1: *septic shock*; 2: *colon perforation leading to stercoral peritonitis*; 3: *Acute Respiratory Distress Syndrome*; 4: *multiple organ failure*; 5: *HBP: High Blood Pressure.*

	line	text	normalized text	ICD codes
Raw	1	choc septique		-
	2	peritonite stercorale sur perforation colique		-
	3	Syndrome de détresse respiratoire aiguë		-
	4	defaillance multivicerale		-
	5	HTA		-
Computed	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	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

Table 2. One sample document from the Hungarian corpus (KSH-HU Death Certificates Corpus). English translations for each *raw* line follow: 1: *respiratory failure*; 3: *bacterial pneumonia*; 4: *pulmonary bronchitis, hepatic metastasis, cerebral metastasis.*

type	line	text	ICD codes
Raw	1	légzési elégt	-
	3	bakt tgy	-
	4	tüdõ hörgõ rd, máj áttét,agy áttét	-
Computed	1		J968
	3		J159
	4		C349
	4		C787
	4		C793

where historical data is available to train systems that can then be applied to current data to assist with new document curation. For French, the training set covered the 2006–2014 period, and the test set from 2015. For Hungarian,

Table 3. One sample document from the Italian corpus (ISTAT-IT Death Certificates Corpus). English translations for each *raw* line follow: 1: *neoplastic cachexia*; 2: *atrial fibrillation with rapid ventricular response*; 3: *cardio-circulatory decompensation, respiratory decompensation*; 4: *pulmonary neoplasia*; 6: *sigmoid resection for neoplasia, COPD (Chronic Obstructive Pulmonary Disease), hypothyroidia.*

type	line	text	ICD codes
Raw	1	CACHESSIA NEOPLASTICA	-
	2	FA AD ELEVATA RISPOSTA VENTRICOLARE	-
	3	SCOMPENSO CARDIOCIRCOLATORIO, SCOMPENSO RESPIRATORIO	-
	4	NEOPLASIA POLMONARE	-
	6	RESEZIONE DEL SIGMA PER NEOPLASIA , BPCO , IPOTIROIDISMO	- -
Computed	1		C809
	2		I489
	2		I471
	3		I516
	3		J988
	4		C349
	6		Y836
	6		D48
	6		J448
	6		E0399

data was only available for the year 2016, but the training and test sets were nonetheless divided chronologically during that year. While the French dataset offers more documents spread over a nine year period, it also reflects changes in the coding rules and practices over the period. In contrast, the Hungarian dataset is smaller but more homogeneous. The Italian dataset was fabricated from de-identified original death certificates to further preserve patient confidentiality.

Table 4. Descriptive statistics of the Death Certificates datasets in French, Hungarian and Italian. Tokens were counted using the linux wc -w command.

	French		**Hungarian**		**Italian**	
	Training	Test	Training	Test	Training	Test
	(2006–2014)	(2015)	(2016)	(2016)	(2016)	(2016)
Certificates	125,384	11,931	84,703	21,176	14,502	3,618
Lines	368,065	34,918	324,266	81,291	49,825	12,602
Tokens	1,250,232	84,091	666,839	167,507	666,839	167,507
Total ICD codes	509,103	48,948	392,020	98,264	60,955	15,789
Unique ICD codes	3,723	1,806	3,124	2,011	1,443	903
Unique unseen ICD codes	-	70	-	202	-	100

Dataset format. In compliance with the World Health Organization (WHO) international standards, death certificates comprise 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.[4] According to WHO recommendations, the completion of both parts is free of any automatic assistance that might influence the certifying physician. The processing of death certificates, including ICD10 coding, is performed independently of physician reporting. In France, Hungary and Italy, coding of death certificates is performed within 18 months of reporting using the IRIS system [13]. 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 (referred as 'raw causes' thereafter) and a file containing the result of ICD code assignment (referred as 'computed causes' thereafter). The 'computed causes' file may contain normalized text that supports the coding decision and can be used in the creation of dictionaries for the purpose of coding assistance. 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. This makes the datasets more suitable for approaching the coding problem as a text classification task at the document level rather than a named entity recognition and normalization task. We have reported separately on the challenges presented by the separation of data into raw and computed causes, and proposed solutions to merge the French data into a single 'aligned' format, relying on the normalized text supplied with the French raw causes [14]. Table 1 presents a sample of French death certificate in 'raw' and 'aligned' format. It illustrates the challenge of alignment with the line 2 in the raw file "péritonite stercorale sur perforation colique" which has to be mapped to line 4 "peritonite stercorale" (code K65.9) and line 5 "perforation colique" (code K63.1) in the computed file.

Data files. Table 5 presents a description of the files that were provided to the participants: training (*train*) files were distributed at the end of February 2018; test files (*test*, with no gold standard) were distributed at test time (at the end of April 2018); and the gold standard for test files (*test+g* in aligned format, *test, computed* in raw format) were disclosed to the participants after the text phase (in May 2018) so that participants could reproduce the performance measures announced by the organizers.

2.2 ICD10 coding task

The coding task consisted of mapping lines in the death certificates to one or more relevant codes from the International Classification of Diseases, tenth revision (ICD10). For the raw datasets, codes were assessed at the certificate level. For the aligned dataset, codes were assessed at the line level.

[4] As can be seen in the sample documents, the line numbering in the raw causes file may (Table 2) or may not (Table 1) be the same in the computed causes file. In some cases, the ordering in the computed causes file was changed to follow the causal chain of events leading to death.

Table 5. Data files. Files after the dashed lines are test files; files after the dotted lines contain the gold test data. L = language (fr = French, hu = Hungarian, it = Italian).

	L.	Split	Type	Year	File name
Aligned	fr	train	aligned	2006–2012	AlignedCauses_2006-2012.csv
	fr	train+g	aligned	2006–2012	AlignedCauses_2006-2012full.csv
	fr	train	aligned	2013	AlignedCauses_2013.csv
	fr	train+g	aligned	2013	AlignedCauses_2013full.csv
	fr	train	aligned	2014	AlignedCauses_2014.csv
	fr	train+g	aligned	2014	AlignedCauses_2014full.csv
	fr	test	aligned	2015	AlignedCauses_2015F_1.csv
	fr	test	list	2015	GoldStandardFR2008_IDs.out
	fr	test+g	aligned	2015	AlignedCauses_2015_full_2018_UTF8_filtered_1m_commonRaw.csv
Raw	fr	train	raw	2006–2012	CausesBrutes_FR_2006–2012.csv
	fr	train	ident	2006–2012	Ident_FR_training.csv
	fr	train+g	computed	2006–2012	CausesCalculees_FR_2006–2012.csv
	fr	train	raw	2013	CausesBrutes_FR_2013.csv
	fr	train	ident	2013	Ident_FR_2013.csv
	fr	train+g	ident	2013	Ident_FR_2013_full.csv
	fr	train+g	computed	2013	CausesCalculees_FR_2013.csv
	fr	train	raw	2014	CausesBrutes_FR_2014.csv
	fr	train	ident	2014	Ident_FR_2014.csv
	fr	train+g	ident	2014	Ident_FR_2014_full.csv
	fr	train+g	computed	2014	CausesCalculees_FR_2014.csv
	fr	test	raw	2015	CausesBrutes_FR_2015F_1.csv
	fr	test	ident	2015	Ident_FR_2015F_1.csv
	fr	test	list	2015	GoldStandardFR2008_IDs.out
	fr	test+g	computed	2015	CausesCalculees_2015_full_2018_UTF8_filtered_1m_commonRaw.csv
Raw	hu	train	raw	2016	CausesBrutes_HU_1.csv
	hu	train	ident	2016	Ident_HU_1.csv
	hu	train+g	computed	2016	CausesCalculees_HU_1.csv
	hu	test	raw	2016	CausesBrutes_HU_2.csv
	hu	test	ident	2016	Ident_HU_2.csv
	hu	test+g	computed	2016	CausesCalculees_HU_2.csv
Raw	it	train	raw	2016	CausesBrutes_IT_1.csv
	it	train	ident	2016	Ident_IT_1.csv
	it	train+g	computed	2016	corpus/CausesCalculees_IT_1.csv
	it	test	raw	2016	CausesBrutes_IT_2.csv
	it	test	ident	2016	Ident_IT_2.csv
	it	test+g	computed	2016	CausesCalculees_IT_2.csv

2.3 Evaluation metrics

System performance was assessed by the usual metrics of information extraction: precision (Formula 1), recall (Formula 2) and F-measure (Formula 3; specifically, we used β=1.).

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{1}$$

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{2}$$

$$\text{F-measure} = \frac{(1+\beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \tag{3}$$

Results were computed using two perl scripts, one for the **raw** datasets (in French, Hungarian and Italian) and one for the **aligned** dataset (in French only). The evaluation tools were supplied to task participants along with the training data. Measures were computed for all causes in the datasets, i.e. the evaluation covered all ICD codes in the test datasets.

For the raw datasets, matches (true positives) were counted for each ICD10 full code supplied that matched the reference for the associated document.

For the aligned dataset, matches (true positives) were counted for each ICD10 full code supplied that matched the reference for the associated document line.

This year, we also experimented with a secondary metric, which consisted in computing recall over the primary causes of death. In death certificate coding, once all the relevant causes of death have been identified in all certificate lines, the chain of events leading to the dealth is analyzed to yield one single *primary cause* of death, which is central to national statistics reporting. This primary cause was available to us for the French and Italian datasets. *Primary recall* was therefore computed as the number of certificates where the primary cause was retrieved by systems over the total number of certificates.

3 Results

Participating teams included between one and nine team members and resided in Algeria (team techno), Canada (team TorontoCL), China (teams ECNU and WebIntelligentLab), France (teams APHP, IAM, ISPED), Germany (team WBI), Italy (Team UNIPD), Spain (teams IxaMed, SINAI and UNED), Switzerland (team SIB) and the United Kingdom (team KCL).

For the Hungarian raw dataset, we received 9 official runs from 5 teams. For the Italian raw dataset, we received 12 official runs from 7 teams. For the French raw dataset, we received 18 official runs from 12 teams. We also received three additional non-official runs from 2 teams, including one run implementing corrections for a faulty official run. For the French aligned dataset, we received 16 official runs from 8 teams. We also received three additional non-official runs from 2 teams, including one run implementing corrections for a faulty official run.

3.1 Methods implemented in the participants' systems

Participants relied on a diverse range approaches including classification methods (often leveraging neural networks), information retrieval techniques and dictionary matching accommodating for different levels of lexical variation. Most participants (12 teams out of 14) used the dictionaries that were supplied as part of the training data as well as other medical terminologies and ontologies (at least one team).

ECNUica. The methods implemented by the ECNUica team [15] combine statistical machine learning and symbolic algorithms together to solve the ICD10 coding task. First they utilize the regular match expressions to mapping test data and find out the ICD10 codes. What's more, in order to handle the data which have no mapping ICD10 codes, they use attributes such as gender and age in the corpus as the feature data to train the random forest and Xgboost model. And then, all the data is classified into A-Z 26 categories, so they use rule-based and similarity computation method to match the classified data with training data. Finally they obtain the specific ICD10 codes of the test data.

ECSTRA-APHP. The ECSTRA-APHP team [16] cast the task as a machine learning problem involving the prediction of the ICD10 codes (categorical variable) from the raw text transformed into word embeddings. We rely on probabilistic convolutional neural network for classification. In the present work, we train a CNN with that uses multiple filters (with varying window sizes) to obtain multiple features on top of word vectors obtained as the first hidden layer of the classification itself. Due to very week representation for the some of ICD codes, we complete prediction with dictionary-based lexical matching classifier which rely on word recognition from a knowledge base build from several available dictionaries on the French ICD 10 classification : second volume of ICD, orphanet thesaurus, French SNOMED CT, and CépiDC dictionaries provided for the challenge.

IAM-ISPED The method used by the IAM ISPED team [17] is a dictionary-based approach. It uses the terms of a terminology (ICD10) to assign ICD10 codes to each text line. The program has a module of typos detection that runs a Levenshtein distance and a module of synonyms expansion (Ins =¿ Insuffisance). The runs1 and 2 differ by the terms used : in run2, all the terms of the column "Standard text" in AlignedCauses files (2006-2012;2013;2014) were used, which corresponded to 42,439 terms and 3,539 codes; in run1, the terms of run2 and the terms in the "Dictionnaire2015.csv" file were used, which corresponded to 148,447 terms and 6,392 codes. The source code of the program will be released.

IMS-UNIPD. Team UNIPD [18] aimed to implement 1) a minimal expert system based on rules to translate acronyms, 2) together with a binary weighting approach to retrieve the items in the dictionary most similar to the portion of the certificate of death, and 3) a basic approach to select the class with the highest weight.

IxaMed. The IxaMed group [19] has approached the automatic ICD10 coding for French, Italian and Hungarian with a neural model that tries to map the input text snippets with the output ICD10 codes. Their solution does not make assumptions about the content of the input and output data, treating them by means of a machine learning approach that assigns a set of labels to any input line. The solution is language-independent, in the sense that treating a new language only needs a set of (input, output) examples, making no use of language-specific information apart from terminological resources such as ICD10 dictionaries, when available.

KCL-Health-NLP. The KCL-Health-NLP team [20] employed a document-level encoder-decoder neural approach. The convolutional encoder operates at the character level. The decoder is recurrent. For French, they contrast the usage of only Raw Text, as well as this text combined with string matched ICD codes. The string matching approach relies on the dictionaries provided, and uses a word n-gram (1-5) representation (ignoring diacritics, including stemming and removal of stopwords) to search for matches. For Italian, they take advantage of language-independent character-level characteristics and contrast results with and without pre-training using the French data. External resources are not used.

LSI-UNED. The LSI-UNED team [21] submitted two runs for each raw dataset. A supervised learning system (run 2) has been implemented using multilayer perceptrons and an One-vs-Rest (OVR) strategy. The training of models was carried out with the training data and dictionaries of CépiDC, estimating the frequency of terms weighted with Bi-Normal Separation (BNS). Additionally, this approach has been supplemented with IR methods in a second system (run 1). To this end, the bias has been limited, generating learning models for the ICD-10 codes that appear more than 100 times in the training dataset. The unclassified diseases by these models are used to build queries and apply them to search engines with code descriptions.

SIB-BITEM. The BITEM-SIB [22] leveraged the large size and textual nature of the training data by investigating an instance-based learning approach. The 360,000 annotated sentences contained in the training data were indexed with a standard search engine. Then, the k-Nearest Neighbors of an input sentence were exploited in order to infer potential codes, thanks to majority voting. A dictionary-based approach was also used for directly mapping codes in sentences, and both approaches were linearly combined.

SINAI. The SINAI team [23] made a system based on Natural Language Processing (NLP) techniques to detect International Classification Diseases (ICD10) codes using different machine learning algorithms. First, their system found all the possibles ICD10 codes looking for how many words of each code exist in the text. Next, several measures of quality of these codes were calculated. With these metrics, different machine learning algorithms were trained and finally the best model was selected to use in the system. Most of the techniques used are independent of the language, therefore the system is easily adaptable to other languages.

KR-ISPED. The SITIS-ISPED team [24] used a deep learning approach and relied on the training data supplied: they used OpenNMT-py, an open source framework for Neural Machine Translation (seq2seq), implemented in PyTorch.

To transform diagnostics into ICD10 codes they utilize an encoder-decoder architecture, consisting of two recurrent neural networks combined together with an attention mechanism. First, the diagnostics and their ICD10 codes are extracted from the csv files and then respectively split into a source text file and a target text file. This extraction is made by a simple bash program. In this way the data consists of parallel source (diagnosis) and target (ICD10 codes) data containing one sentence per line with words separated by a space. Then those data are split into two groups: one for training and one for validation. Validation files are used to evaluate the convergence of the training process. For source files, a first preprocessing step converts upper cases into lower cases. A tokenization process is applied on sources files and on target files which are used as input for the neural network The used encoder/decoder model consists of a 2 layers LSTM with 500 hidden units on both the encoder and decoder. The encoder encodes the input sequence into a context vector which is used by the decoder to generate the output sequence. The training process goes on for 13 epochs and provide a model. From the test data provided by the CLEF organization, we extracted the diagnostics, preprocessed them and used the model we created to "translate" them into their respective ICD10 codes.

Techno. The techno team [25] developed Naive Bayes (NB) classifier for text classification to information extraction from written text at CLEF eHealth 2018 challenge, task1. We used a NB classifier to generate a classification model. The evaluation of the proposed approach does not show good performance.

TorontoCL. The TorontoCL team [26] assigned ICD-10 codes to cause-of-death phrases in multiple languages by creating rich and relevant word embedding models. They train 100-dimensional word embeddings on the training data provided, as well as on language-specific Wikipedia corpora. they then use an ensemble model for ICD coding prediction which includes n-gram matching of the raw text to the provided ICD dictionary followed by an ensemble of a convolutional neural network and a recurrent neural network encoder-decoder.

WBI. The contribution of the WBI team [11] focus on the setup and evaluation of a baseline language-independent neural architecture as well as a simple, heuristic multi-language word embedding space. Their approach builds on two recurrent neural networks models and models the extraction and classification of death causes as two-step process. First, they employ a LSTM-based sequence-to-sequence model to obtain a death cause from each death certificate line. Afterwards, a bidirectional LSTM model with attention mechanism will be utilized to assign the respective ICD-10 codes to the received death cause description. Both models represent words using pre-trained fastText word embeddings. Independently from the original language of a word they represent it by looking up the word in the embedding models of the three languages and concatenate the obtained vectors to build heuristic shared vector space.

WebIntelligentLab. The WebIntelligentLab team used a deep learning method viz. lstm with fully connected layers that uses only training data, no dictionary, and other external data.

Baseline. To provide a better assessment of the task difficulty and system performance, this year we offered results from a so-called *frequency baseline*, which consisted in assigning to a certificate line from the test set the top 2 most frequently associated ICD10 codes in the training and development sets, using case and diacritic insensitive line matching.

3.2 System performance on death certificate coding

Tables 6 to 9 present system performance on the ICD10 coding task for each dataset. Team IxaMed obtained the best performance in terms of F-measure for all datasets. However, we can note that the overall recall perfromance did not always align with the recall computed over primary causes of death (for French and Italian only).

4 Discussion

In this section, we discuss system performance as well as dataset composition and we highlight directions for future work.

4.1 Natural Language Processing for assisting death certificates coding

System performance generally far exceeded the baseline for all three languages. The best systems achieved high precision (.846 F-measure and above) as well as high recall (.597 for French, .955 for Hungarian and .945 for Italian). Similarly to last year, we observe a gap in recall performance between the raw and aligned version of the French dataset, which suggests that there is value in performing the line alignment of the training data. We also note that the primary cause of death recall is higher on the aligned vs. raw format. Many systems offered higher primary cause of death recall than overall recall on the aligned dataset.

Although no direct comparison is possible because the test sets were different, we can notice that the best performance from last year (.825 F-mesure for French raw, .867 F-mesure for French aligned by the LIMSI team [27]) remains ahead of this year's achievements.

The results of the submitting systems show consistent performance across languages for those that addressed more than one language. Of note, all systems but one set up a common architecture for the different languages, that then independently leveraged the resources available in each language (i.e. pre-processing, training corpus, dictionaries, external corpora used to create word embeddings...) Only one team [11] attempted to develop a unique system that could address all three languages, with varying success depending on the language. They also

Table 6. System performance for ICD10 coding on the **French aligned** test corpus in terms of Precision (P), recall (R) and F-measure (F). A horizontal dash line places the frequency baseline performance. The top part of the table displays official runs, while the bottom part displays non-official and baseline runs.

	French (Aligned)				
	Team	P	R	F	Primary R
Official runs	IxaMed-run2	.841	**.835**	**.838**	**.819**
	IxaMed-run1	**.846**	.822	.834	.814
	IAM-run2	.794	.779	.786	.770
	IAM-run1	.782	.772	.777	.757
	SIB-TM	.763	.764	.764	.777
	TorontoCL-run2	.810	.720	.762	.702
	TorontoCL-run1	.815	.712	.760	.694
	KCL-Health-NLP-run1	.787	.553	.649	.629
	KCL-Health-NLP-run2	.769	.537	.632	.621
	SINAI-run2	.733	.534	.618	.549
	SINAI-run1	.725	.528	.611	.527
	WebIntelligentLab	.673	.491	.567	.451
	ECNUica-run1	.771	.437	.558	.526
	ECNUica-run2	.771	.437	.558	.526
	techno	.489	.356	.412	.410
	KR-ISPED	.029	.020	.023	.029
	average	.712	.581	.634	.589
	median	.771	.545	.641	.621
Non-off.	APHP-run1	.634	.600	.621	.653
	APHP-run2	.794	.607	.688	.713
	KR-ISPED-corrected	.665	.453	.539	.524
	Frequency baseline	.452	.450	.451	.495

report that their method still has room for improvement as it currently handles the task as a classification method that assigns one and only one code per death certificate line, which significantly limits the recall performance.

Overall, the level of performance achieved by participants this year shows great potential for assisting death certificate coders throughout Europe in their daily task.

4.2 Limitations

Size of the French test set. The French test set initially distributed this year comprised 24,375 death certificates in the raw and aligned format. Owing to a bug in the selection process, only 11,931 certificates were present in both raw and aligned format. In order to make the results directly comparable between formats, system performance was eventually computed on the subset of 11,931 common certificates. Even though the final size of the test is smaller than initially planned, we believe that the test set is still large enough to provide interesting insight on system performance for death certificate coding in French.

Table 7. System performance for ICD10 coding on the **French raw** test corpus in terms of Precision (P), recall (R), F-measure (F) and recall on Primary Cause of Death (Primary R) A horizontal dash line places the frequency baseline performance. The top part of the table displays official runs, while the bottom part displays non-official and baseline runs.

	French (Raw)				
	Team	P	R	F	Primary R
Official runs	IxaMed-run1	.872	**.597**	**.709**	.579
	IxaMed-run2	.877	.588	.704	.573
	LSI-UNED-run1	.842	.556	.670	.535
	LSI-UNED-run2	.879	.540	.669	.506
	IAM-run2	.820	.560	.666	.555
	IAM-run1	.807	.555	.657	.544
	TorontoCL-run2	.842	.522	.644	.507
	TorontoCL-run1	.847	.515	.641	.500
	WebIntelligentLab	.702	.495	.580	.451
	ECNUica-run1	.790	.456	.578	.530
	KCL-Health-NLP-run1	.738	.405	.523	.430
	KCL-Health-NLP-run2	.724	.394	.510	.421
	ims-unipd	.653	.396	.493	.401
	techno	.569	.286	.380	.349
	WBI-run2	.512	.253	.339	.302
	WBI-run1	.494	.246	.329	.293
	KR-ISPED	.043	.021	.028	.015
	ECNUica-run2	**1.000**	0.000	.000	.000
	average	.723	.410	.507	.414
	median	.798	.475	.579	.500
Non-off.	APHP-run1	.668	.601	.633	.613
	APHP-run2	.816	.607	.696	**.713**
	KR-ISPED-corrected	.676	.323	.437	.377
	Frequency baseline	.341	.201	.253	.221

Comparability across languages. Overall system performance seem to be higher on the Hungarian (average F-measure .80) and Italian (average F-measure .799) datasets, compared to French (raw average F-measure .507). However, the question of strict comparability across languages remains open because of the differences in nature between the datasets. The Italian dataset is a synthetic dataset fabricated using selected real data. It is possible that the selection process yielded somewhat content that was more standard and more easy to analyze in order to reach the consistency goals for the final synthetic certificates. The Hungarian dataset was obtained from transcribed paper certificates. It is possible that some of the natural language difficulties present in the original paper certificates (such as typos) were smoothed out during the transcription process, which was performed manually by contractors. The French dataset was obtained directly from electronic certification, which means that it contains the original text exactly as entered by doctors without any filtering of difficulties. The prac-

Table 8. System performance for ICD10 coding on the **Hungarian raw** test corpus in terms of Precision (P), recall (R) and F-measure (F).

Hungarian (Raw)			
Team	P	R	F
IxaMed run2	**.970**	**.955**	**.963**
IxaMed run1	.968	.954	.961
LSI UNED-run2	.946	.911	.928
LSI UNED-run1	.932	.922	.927
TorontoCL-run2	.922	.897	.910
TorontoCL-run1	.901	.887	.894
ims unipd	.761	.748	.755
WBI-run2	.522	.388	.445
WBI-run1	.518	.384	.441
average	.827	.783	.803
median	.922	.897	.910
Frequency baseline	.243	.174	.202

Table 9. System performance for ICD10 coding on the **Italian raw** test corpus in terms of Precision (P), recall (R), F-measure (F) and primary cause of death recall (Primary R).

Italian (Raw)				
Team	P	R	F	Primary R
IxaMed run1	**.960**	**.945**	**.952**	.705
IxaMed run2	.945	.922	.934	.699
LSI UNED-run1	.917	.875	.895	.666
LSI UNED-run2	.931	.861	.895	.616
TorontoCL-run1	.908	.824	.864	.650
TorontoCL-run2	.900	.829	.863	.652
WBI-run2	.862	.689	.766	**.715**
WBI-run1	.857	.685	.761	.712
KCL-Health-NLP-run1	.746	.636	.687	.492
KCL-Health-NLP-run2	.725	.616	.666	.492
ims unipd	.535	.484	.509	.375
average	.844	.761	.799	.616
median	.900	.824	.863	.652
Frequency baseline	.165	.172	.169	.071

tice of writing death certificates in the three different countries may also generate notable differences in the writing style or depth of descriptions that impact the analysis. A further exploration of dataset characteristics in terms of number of typos, acronyms or token/type ratios could yield interesting insight on the comparability of data across languages.

5 Conclusion

We released a new set of death certificates to evaluate systems on the task of ICD10 coding in multiple languages. This is the fourth edition of a biomedical NLP challenge that provides large gold-standard annotated corpora in a language other than English. Results show that high performance can be achieved by NLP systems on the task of coding for death certificates in French, Hungarian and Italian. The level of performance observed shows that there is potential for integrating automated assistance in the death certificate coding workflow in all three languages. The corpus used and the participating team system results are an important contribution to the research community. The comparable corpora could be used for studies that go beyond the scope of the challenge, including a cross-country analysis of death certificate contents. In addition, the focus on three languages other than English (French, Hungarian and Italian) remains a rare initiative in the biomedical NLP community.

Acknowledgements

We want to thank all participating teams for their effort in addressing new and challenging tasks. The organization work for CLEF eHealth 2018 task 1 was supported by the Agence Nationale pour la Recherche (French National Research Agency) under grant number ANR-13-JCJC-SIMI2-CABeRneT. The CLEF eHealth 2018 evaluation lab has been supported in part by the CLEF Initiative and Data61.

References

1. Suominen, H., Kelly, L., Goeuriot, L., Kanoulas, E., Azzopardi, L., Spijker, R., Li, D., Névéol, A., Ramadier, L., Robert, A., Zuccon, G., Palotti, J. Overview of the CLEF eHealth Evaluation Lab 2018. In: CLEF 2018 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September. (2018)8
2. World Health Organization. ICD-10. International Statistical Classification of Diseases and Related Health Problems. 10th Revision. Volume 2. Instruction manual. 2011.
3. Névéol, A., Anderson, R.N., Cohen, K.B., Grouin, C., Lavergne, T., Rey, G., Robert, A., Zweigenbaum, P.: CLEF eHealth 2017 multilingual information extraction task overview: Icd10 coding of death certificates in english and french. In: CLEF 2017 Online Working Notes. CEUR-WS (2017)
4. Suominen H, Salantera S, Velupillai S, Chapman WW, Savova G, Elhadad N, Pradhan S, South BR, Mowery DL, Jones GJF, Leveling J, Kelly L, Goeuriot L, Martinez D, Zuccon G. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Forner P, Müller H, Paredes R, Rosso P, Stein B (eds), Information Access Evaluation. Multilinguality, Multimodality, and Visualization. LNCS (vol. 8138):212-231. Springer, 2013

5. Goeuriot L, Kelly L, Suominen H, Hanlen L, Névéol A, Grouin C, Palotti J, Zuccon G. Overview of the CLEF eHealth Evaluation Lab 2015. In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction. Springer, 2015
6. Kelly L, Goeuriot L, Suominen H, Névéol A, Palotti J, Zuccon G. (2016) Overview of the CLEF eHealth Evaluation Lab 2016. In: Fuhr N. et al. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2016. Lecture Notes in Computer Science, vol 9822. Springer, Cham
7. Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, João Palotti, and Guido Zuccon. CLEF 2017 eHealth Evaluation Lab Overview. CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer, September, 2017.
8. Kelly L, Goeuriot L, Suominen H, Schreck T, Leroy G, Mowery DL, Velupillai S, Chapman WW, Martinez D, Zuccon G, Palotti J. Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In: Kanoulas E, Lupu M, Clough P, Sanderson M, Hall M, Hanbury A, Toms E (eds), Information Access Evaluation. Multilinguality, Multimodality, and Interaction. LNCS (vol. 8685):172-191. Springer, 2014
9. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc, 18(5):540-3
10. Huang CC, Lu Z (2015). Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform, 2015 May 1. pii: bbv024.
11. Ševa J, Sänger M, and Leser U (2018). WBI at CLEF eHealth 2018 Task 1: Language-independent ICD-10 coding using multi-lingual embeddings and recurrent neural networks. CLEF 2018 Online Working Notes. CEUR-WS
12. Pavillon G., Laurent F (2003). Certification et codification des causes médicales de décès. Bulletin Epidémiologique Hebdomadaire - BEH:134-138. `http://opac.invs.sante.fr/doc_num.php?explnum_id=2065` (accessed: 2016-06-06)
13. Johansson LA, Pavillon G (2005). IRIS: A language-independent coding system based on the NCHS system MMDS. In WHO-FIC Network Meeting, Tokyo, Japan
14. Lavergne T, Névéol A, Robert A, Grouin C, Rey G, Zweigenbaum P. A Dataset for ICD-10 Coding of Death Certificates: Creation and Usage. Proceedings of the Fifth Workshop on Building and Evaluating Ressources for Health and Biomedical Text Processing - BioTxtM2016. 2016.
15. Li M, Xu C, Wei T, Bao D, Lu N, and Yang J (2018). ECNU at 2018 eHealth Task1 Multilingual Information Extraction. CLEF 2018 Online Working Notes. CEUR-WS
16. Flicoteaux R (2018). ECSTRA-APHP @ CLEF eHealth2018-task 1: ICD10 Code Extraction from Death Certificates. CLEF 2018 Online Working Notes. CEUR-WS
17. Cossin S, Jouhet V, Mougin F, Diallo G, and Thiessard F (2018). IAM at CLEF eHealth 2018 : Concept Annotation and Coding in French Death Certificates. CLEF 2018 Online Working Notes. CEUR-WS
18. Di Nunzio GM (2018). Classification of ICD10 Codes with no Resources but Reproducible Code. IMS Unipd at CLEF eHealth Task 1. CLEF 2018 Online Working Notes. CEUR-WS
19. Atutxa A, Casillas A, Ezeiza N, Goenaga I, Fresno V, Gojenola K, Martinez R, Oronoz M and Perez-de-Viñaspre O (2018). IxaMed at CLEF eHealth 2018 Task 1: ICD10 Coding with a Sequence-to-Sequence approach. CLEF 2018 Online Working Notes. CEUR-WS

20. Ive J, Viani N, Chandran D, Bittar A, and Velupillai S (2018). KCL-Health-NLP@CLEF eHealth 2018 Task 1: ICD-10 Coding of French and Italian Death Certificates with Character-Level Convolutional Neural Networks CLEF 2018 Online Working Notes. CEUR-WS
21. Almagro M, Montalvo S, Diaz de Ilarraza A, and Pérez A (2018). LSI UNED at CLEF eHealth 2018: A Combination of Information Retrieval Techniques and Neural Networks for ICD-10 Coding of Death Certificates. CLEF 2018 Online Working Notes. CEUR-WS
22. Gobeill J and Ruch P (2018). Instance-based learning for ICD10 categorization. CLEF 2018 Online Working Notes. CEUR-WS
23. Lopez-Úbeda P, Diaz-Galiano MC, Martin-Valdivia MT, and Ureña-López LA (2018). Machine learning to detect ICD10 codes in causes of death. CLEF 2018 Online Working Notes. CEUR-WS
24. Réby K, Cossin S, Bordea G, and Diallo G (2018). SITIS-ISPED in CLEF eHealth 2018 Task 1: ICD10 coding using Deep Learning. CLEF 2018 Online Working Notes. CEUR-WS
25. Bounaama R and El Amine Abderrahim M (2018). Tlemcen University at CELF eHealth 2018 Team techno: Multilingual Information Extraction - ICD10 coding. CLEF 2018 Online Working Notes. CEUR-WS
26. Jeblee S, Budhkar A, Milić S, Pinto J, Pou-Prom C, Vishnubhotla K, Hirst G, and Rudzicz F (2018). TorontoCL at the CLEF 2018 eHealth Challenge Task 1. CLEF 2018 Online Working Notes. CEUR-WS
27. Zweigenbaum P and Lavergne T (2017). Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates. CLEF 2017 Online Working Notes. CEUR-WS