LITL at CLEF eHealth2017: automatic classication of death reports

Lydia-Mai Ho-Dac, Cécile Fabre, Anouk Birski, Imane Boudraa, Aline Bourriot, Manon Cassier, Léa Delvenne, Charline Garcia-Gonzalez, Eun-Bee Kang, Elisa Piccinini, et al.

LITL at CLEF eHealth2017: automatic classification of death reports

Lydia-Mai Ho-Dac[1], Cécile Fabre[1], Anouk Birski[2], Imane Boudraa[2], Aline Bourriot[2], Manon Cassier[2], Léa Delvenne[2], Charline Garcia-Gonzalez[2], Eun-Bee Kang[2], Elisa Piccinini[2], Camille Rohrbacher[2], and Aure Séguier[2]

[1] CLLE, University of Toulouse, CNRS, UT2J, France
{hodac,cecile.fabre}@univ-tlse2.fr,
WWW home page: http://clle.univ-tlse2.fr/accueil/equipe-erss/cartel/
[2] Master LITL, University of Toulouse, UT2J, France

Abstract. This paper describes the participation of a group of students supervised by two teachers to the CLEF eHealth 2017 campaign, task 1. The task involves the classification of death certificates in French and more precisely the labelling of each cause of death with the relevant ICD10 code. The system that performs the automatic coding is based on an information retrieval method using the Solr interface. Two runs were submitted according to whether the system distinguishes cases of multiple causes or not. The best performance was obtained with the system which distinguishes multiple causes, with a precision of 0.61 and a recall of 0.55.

Keywords: Text classification, biomedical texts, code assignment, cause of death extraction, information retrieval

1 Introduction

This article describes the participation of a group of master's students to the CLEF eHealth 2017[4] Lab which aims at gathering research on NLP techniques dedicated to improve information retrieval and extraction in biomedical texts.

LITL (*Linguistique, Informatique, Technologies du Langage*, i.e. Linguistics, IT, Language technologies) is a master's program at the University of Toulouse, France. Mainly dedicated to students coming from a linguistics and humanities background, it comprises, for a major part, courses in natural language processing (NLP), computational linguistics and practical aspects of corpus analysis mixing programming and the use of various tools. An important part of this curriculum is project-oriented, and the first year students have to build a fully operational processing system for a precise NLP task. Last year, the students participated to the CLEF eHealth challenge task 2 which consisted in the recognition and categorization of medical entities in French biomedical documents [5]. This year's project was the participation to the CLEF eHealth challenge, more precisely task 1: multilingual Information Extraction [9] which concerns the automatic coding of death certificates.

The supervisors of this project considered this task ideal for pedagogical purposes for the following reasons:

- information retrieval and extraction is a well-known, well-defined and central task in modern NLP;
- the biomedical domain gives access to linguistic resources and data which perfectly illustrate applied linguistics, with different degrees of normalization ranging from raw natural language data to codification via standardized text;
- a collaborative task is an excellent exercise for students, as it motivates them and gives them a clear feedback on their work;
- the target language of CLEF eHealth 2017 is French, which is the students' working language;
- the task's schedule was ideally suited to the master's calendar.

Working as a team along the entire semester, the students were thus able (with help from their teachers) to submit two runs for the selected task, and got very satisfactory results for a first attempt.

This paper is organized as follows. Sect. 2 describes the task and get a closer look at the data. Sects. 3 and 4 give a precise description of the different components of the system designed for the task, while the results are given and discussed in Sect. 5.

2 Task Description

2.1 Overview

The CLEF 2017 eHealth task 1 consists of mapping statements in the death certificates to one or more relevant codes from the International Classification of Diseases, tenth revision (ICD10). Death certificates are mandatory documents written by medical practitioners after any death occurring on a territory (the French territory for French data). The systematic coding of causes of death is essential for monitoring public health, providing data for the quantitative analyses and international comparisons of epidemiological data [11]. The Word Health Organization (WHO) manages and provides multilingual resources for standardizing the coding task. According to the WHO international standards, the terminology that has to be applied for causes of death coding is the ICD10. Unfortunately, death certificates are still mainly hand-written and the ICD10 code is not assigned by the practitioner at the time of the death statement [6].

In such a context, computer-assisted coding tools are required for facilitating and speeding up the coding process [11]. Automatic coding in the ICD has been the subject of a number of studies for English (e.g. [16, 12, 8]) but few and only recently for French[3] [1, 17, 14, 3, 7, 2].

Two steps were distinguished for automatic coding: automatically detecting single causes of death statements then categorizing each detected statement according to the ICD10 taxonomy. Before describing these two steps (3), the next sections present the data and the ICD10 taxonomy.

[3] Thanks to the CLEF eHealth challenge.

2.2 Data: the CépiDC Causes of Death Corpus

The training, development and test data sets are part of the CépiDC (Epidemiology Center on medical causes of death) Causes of Death Corpus provided by the INSERM, the French Institute for Health and Medical Research [11]. The death certificates composing the CépiDC corpus were collected from physicians and hospitals in France over the period of 2006–2014. They correspond to forms where practitioners record various data, as shown in Fig. 1.

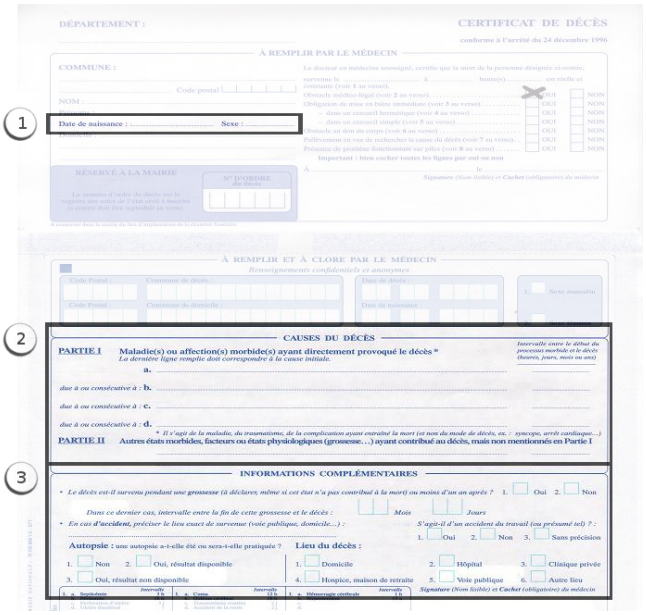

Fig. 1. French death certificate form

The recorded data provided in the CépiDC Corpus are:

- age and gender of the deceased (① in Fig. 1),
- information about death circumstances ③ and
- causes of death ② with a distinction between direct causes (the 4 top lines a,b,c,d, in ②) and indirect causes (the bottom line in ②, 'PARTIE II') of the death.

Example 1 is extracted from the test data set. It gives an excerpt of one certificate as encoded in the CépiDC corpus with, from left to right: the death certificate ID (e.g. *100029*), its year of processing (e.g. *2014*), the gender of the deceased (*1* or *2*), the age at the time of death (e.g. *60*), the location of death (e.g. *2* for hospital), the line number within the death certificate (ideally, each line numbered from 1 to 4 must correspond to one single direct cause, while line

5 contains all the indirect causes) and the raw text contained in the line. The two last fields provide information about how long the patient had been suffering from coded cause (*2;23* means 23 hours, *5;2* means 2 years).

Example 1.

100029;2014;1;60;2;1;altération de l'état général;3;23
100029;2014;1;60;2;2;tumeur de l'épiglotte du larynx étendue métastases [...];3;23
100029;2014;1;60;2;5;tabagisme chronique;NULL;NULL

100055;2014;2;75;2;1;syndrome de defaillance multiviscerale;3;2
100055;2014;2;75;2;2;choc septique;3;26
100055;2014;2;75;2;3;pylonephrite aiguë;3;26
100055;2014;2;75;2;5;HTA, dyslipidemie, diabete type 2, ;5;0

100056;2014;1;50;2;1;CANCER BRONCHIQUE AVEC metastases [...];NULL;NULL
100056;2014;1;50;2;5;etat cachectique;NULL;NULL

100089;2014;2;45;2;1;CANCER OVARIEN;5;2

As Example 1 illustrates, texts are written in plain language with a lot of variation in terms of :

- document size (certificate 100055 contains 4 statements whereas certificate 100089 contains 1);
- statement size (*choc septique* vs. *tumeur de l'épiglotte du larynx étendue métastases ganglionnaires et hépatique*);
- graphical norms: case (*CANCER BRONCHIQUE AVEC metastases multiples*) and diacritics (*pylonephrite aiguë*);
- a more or less telegraphic style (*HTA, dyslipidemie, diabete type 2*).

In addition, as domain-specific texts, raw texts contain a lot of biomedical entities (*pylonephrite*, *dyslipidemie*, *cachectique*), acronyms (*HTA*) and abbreviations.

One of the task implied by IDC-10 manual coding is to standardize lexical variants in order to assign the right ICD10 code. As stated in the overview of the corresponding CLEF eHealth 2016 task "While some of the text lines were short and contained a term that could be directly linked to a single ICD10 code, other lines could be run-on"[10]. Such variations may be observed in the training data set where statements are aligned with the gold standard ICD10 code and a descriptor called 'standard text' which corresponds either to the ICD10 code label or to an excerpt of the raw text that supports the selection of an ICD10 code [6]. All available standard texts and their corresponding ICD10 code are provided by the organizers in files called 'Dictionaries'.

As illustrated in Example 2, standard texts (in bold) may be either an abbreviation of the raw text (e.g. *sdra*); a simplification by removing functional words (e.g. *par*, *à*) or expansions (e.g. *sévère*); or conversely an addition to the raw text (e.g. *arme feu* added to the rawText *Suicide* in certificate 72).

Example 2.

syndrome de detresse respitaoire aigu;NULL;NULL;1-1;**sdra**;J80
plaie par arme à feu;NULL;NULL;4-1;**plaie arme feu**;Y249
maladied 'alzheimer sévère;NULL;NULL;3-1;**maladie alzheimer**;G309
Suicide;NULL;NULL;2-1;**suicide arme feu**;X749

The last specificity of CépiDC data is that one line in a certificate may express more than one cause of death and as a consequence corresponds to multiple codes. In the CépiDC corpus, such statements are repeated on as many lines as there are codes, as in Example 3 where the 5th line of the certificate number 11 'Pneumopathie, ethylisme chronique, stéatose hépatique' expresses 3 causes of death aligned with 3 ICD10 codes (J189, K760, F102) and 3 standard texts.

Example 3.
Pneumopathie, ethylisme chronique, stéatose hépatique;...;pneumopathie;J189
Pneumopathie, ethylisme chronique, stéatose hépatique;...;stéatose hépatique;K760
Pneumopathie, ethylisme chronique, stéatose hépatique;...;éthylisme chronique;F102

Lines that are assigned multiple causes often occur in the 5th position which corresponds to all indirect causes involved in the death (in the training data set, half of those lines are linked to more than one cause); nevertheless, the other lines may also be in this case (e.g. around 15% of the 1st lines).
Table 1 gives a quantitative overview of the data sets made available for the task. #statements gives the number of lines in the certificate, #lines corresponds to the number of lines in the aligned data set[4], and #multiple-codes st. gives the amount of lines associated with multiple codes.

Table 1. Quantitative overview of the aligned data sets extracted from the CépiDC Causes of Death Corpus.

data set	**period**	**#certificates**	**#statements**	**#words(rawText)**
training	2006–2012	65,843	195,204	1,176,993
development	2013	27,850	80,899	496,649
test	2014	31,690	91,962	316,855
total	2006–2014	125,383	368,065	1,990,497

data set	**#statements**	**#lines**	**#multiple-codes st.**
training	195,204	266,807	45,387
development	80,899	110,869	18,718
test	91,962	131,426	24,960
total	368,065	509,102	89,065

2.3 ICD10 Taxonomy

The ICD10 taxonomy is the 10th version of the standard list of causes of death which has been adopted by the WHO [5].

[4] The training and development data sets are provided in the aligned format and the Test data set, in two formats: raw and aligned ([6]).
[5] The first edition was adopted by the International Statistical Institute in 1893, see http://www.who.int/classifications/icd/en/.

The overall ICD10 taxonomy contains 40,519 ICD10 codes made up of one letter and a sequence of 2 to 5 digits (e.g. A00, Z04800). Letters refer to a chapter of the taxonomy (e.g. A–B refer to the chapter about infection and parasitic diseases) and digits give access to the hierarchy of the taxonomy (see Table 2).

Table 2. ICD10 hierarchy

code	Description
J	Chapter 10: Diseases of the respiratory system
J96	Respiratory failure, not elsewhere classified
J960	Acute Respiratory failure
J9600	Acute Respiratory failure type I [hypoxique]
J9601	Acute Respiratory failure type II [hypercapnique]
J961	Chronic Respiratory failure
J961+0	Chronic Obstructive Respiratory failure
J96100	Chronic Obstructive Respiratory failure type I [hypoxique]
J96190	Chronic Obstructive Respiratory failure, unspecified
...	...

The code required for the CLEF eHealth2017 task contains 3 or 4 characters, i.e. one letter and 2 or 3 digits[6]. There are 3,233 different codes in the training set, 2,363 in the development set and 2,527 in the gold standard test set.

The code distribution is unbalanced. Over half of the lines are aligned with a code beginning with the letter C, I, J or R which correspond to the chapters about 'Malignant neoplasms', 'Diseases of the circulatory system', 'Diseases of the respiratory system' and 'Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified' respectively[7]. The most frequently assigned code is R092 (*Arrêt cardio respiratoire* i.e. 'cardiorespiratory failure') occurring 9,619, 4,110 and 4,748 times in the training, development, gold standard test sets, far ahead of A419 ('Sepsis, unspecified'), the second most frequent code (6,066+2,634+3,366 times).

The next two sections describe the system we proposed for automatically assigning an ICD10 code to each detected cause of death occurring in the certificates. This task is related to text classification and more specifically to automatic report classification in a specific domain [13]. The cornerstone of the system proposed by the LITL team is the Solr IR toolkit which is used as a search engine for indexing the training set and for querying each cause of death.

The indexing of the training data is detailed in the next section; Sect. 4 describes the transformation of the statements from the test set into queries.

[6] 14,425 ICD10 codes of the overall ICD10 taxonomy contain 3 or 4 characters; the remaining 26,108 have at least 5 characters

[7] 62% in the training set, 64% in the development set, 65% in the gold standard data set; 57% in the run1 we submitted and 60% in the run2.

3 IR System Description and Solr Configuration

As summed up in [10], the systems that took part in the first edition of this shared task opted for different methods, namely: dictionary-based pattern matching, machine learning (including topic modeling) and information retrieval methods. Most of them made use of the lexical resources available in the biomedical field (terminologies and ontologies). They all showed the necessity to handle lexical variation by performing a range of pre-processing operations (taking into account graphical, lexical, and semantic variation). Taking advantage of these previous experiments, we decided to give a special attention to the preprocessing steps and to follow the general idea implemented by [17] to use IR style methods. But while this system targeted only statements with one code, putting aside the association between a single statement and several causes of death, we decided to specifically address this problem by introducing cause splitting.

3.1 Indexing the Training Data and Structuring the Collection

IR systems require to build then index the collection of texts. In line with the LIMSI submission to last year's competition [17], we decided to build a collection where each document corresponds to one single ICD10 code and contains a concatenated version of different types of information available in the training set in association with the ICD10 code. Only ICD10 codes occurring in the training data set are taken into account. ICD10 codes without raw texts are removed from the collection. As a result, our collection is made up of 3,232 documents (corresponding to as many ICD10 codes).

Three collections were tested:

- a minimalist collection with only concatenated raw texts;
- a collection compiling the available metadata in addition to raw texts, i.e. age and gender of the deceased and information about death circumstances, after transforming them into text (*man*, *woman*, *20–30 years*, *hospital*);
- two collections distinguishing raw texts recorded in the first 4 lines (i.e. expressing the direct causes of the death) from raw texts recorded in the 5th line (i.e. the indirect causes).

Because the two latter showed lower performances, we selected the minimalist collection and decided to add features from external resources.

3.2 Adding Features from External Resources: CépiDC Dictionaries and SNOMED Terminology

In order to improve recall, we added to the concatenated raw texts different types of lexical variants extracted from external resources in the biomedical domain. First, we systematically added the description of the code as stated in the ICD taxonomy (see Table 2). Secondly, we extracted terms from the CépiDC dictionaries and the SNOMED resource.

The CépiDC dictionaries, provided as part of the training data, list all the 'diagnostic texts' i.e. standard texts annotated in the CépiDC corpus from 2014 to 2016 with their assigned ICD10 code[8]. These lexicons were manually curated and provide a large amount of lexical variants (163,557 different entries). For example, 381 terms are associated with the code R092 ('Respiratory arrest') such as *respiratory failure*, *cardiac arrest*, *sudden cardiac arrest*, *SCA*, *sudden cardiac death*, *SCD*, etc. **3,085 documents** were enriched with all diagnostic texts extracted from the CépiDC dictionaries.

SNOMED [15] is a well-known and widely used resource for biomedical NLP. It contains extensive word lists, it is multilingual and can easily be mapped with other coding systems, such as ICD10. The 'Diagnostics' category of the French SNOMED contains 42,921 terms among which 34,073 are explicitly linked to an ICD10 code. We added to our collection all the terms with an exact match to the ICD10 code, excluding the SNOMED terms associated with an ICD10 code made up of more than 4 characters (2,714 entries). 2,074 documents were enriched with all relevant SNOMED terms.

As a result, the collection used in the Solr search engine comprises 3,232 documents composed of 5 types of elements: the ICD10 code, the ICD10 description, all the raw texts, all the diagnostic texts and the relevant SNOMED entities. Example 4 extracted from our collection shows how SNOMED complete the ICD heading and CépiDC dictionaries, especially by adding *malaria* to the document relative to *paludism* (the French name for malaria).

Example 4.
```
<field name="id">105</field>
<field name="code">b54</field>
<field name="icd">paludisme, sans précision</field>
<field name="diagText">accès palustre</field>
<field name="diagText">accès pseudo-palustre</field>
<field name="diagText">antécédent crise paludisme</field>
<field name="diagText">choc septique paludéen</field>
<field name="diagText">crise paroxystique paludisme</field>
<field name="diagText">paludisme</field>
<field name="diagText">paludisme chronique</field>
<field name="diagText">paludisme multiviscéral</field>
<field name="diagText">paludisme viscéral</field>
<field name="diagText">paludisme viscéral évolutif</field>
<field name="snomed">paludisme</field>
<field name="snomed">malaria</field>
<field name="snomed">paludisme algide</field>
<field name="snomed">hépatite palustre</field>
<field name="snomed">dépôts de pigments malariens</field>
<field name="snomed">présence de pigments paludéens</field>
<field name="rawText">paludisme</field>
```

3.3 Indexing the Collection

The collection was then indexed by using the following basic tools provided with Solr:

- standard tokenization of text fields,

[8] The other features made available in the dictionaries were not exploited by our system

- lowercasing text fields,
- elision filtering,
- stopwords filtering (using the default Solr stopword list for French),
- light stemming using the French snowball stemmer.

4 Querying Solr

Our information retrieval process is divided into five steps:

1. preprocessing and formatting each line of the test data set;
2. transforming the formatted lines into queries;
3. expanding queries;
4. sending the queries to the Solr search engine;
5. collecting the first document which matches.

The next subsections present our choices regarding data preprocessing and query expansion. These choices are based on the qualitative analysis of a large number of CépiDC lines which led to conclusions which were fairly supported by previous works on the same task [10] and more generally by research on information extraction in a specific domain such as biomedicine:

- texts must be spellchecked because of the large amount of typing errors such as characters switching, e.g. *c**ra**dio* instead of *c**ar**dio*;
- compound words separators must be normalized, since the same word may be written with an hyphen (*cardio-respiratoire*), a space (*cardio respiratoire*) or without a separator (*cardiorespiratoire*);
- abbreviations (*cardio-respi*) and acronyms (*CR*) must be associated to their full forms (*cardio-respiratoire*);
- specialized terms must be associated to their semantic and morphological variants, such as synonyms and inflected or derived forms (*cardio*, *cardiaque*, *coeur*).

These preliminary observations led us also to conclude that statements involving more than one cause of death should be segmented with one cause per line.

4.1 Preprocessing of Statements

Error Correction in the Test Set An error in the aligned data set comes up from a quick look to the test set. In some cases, the raw text contains one or more semicolons which are also used for separating values. As a consequence, values are misplaced as illustrated in examples in 5.

Example 5.
140646;2014;1;35;6;5;HTA traitée ; tabagisme;NULL
141050;2014;1;80;2;5;état grabataire ; démence mixte; DNID
142741;2014;2;55;2;5;Sepsis à hafnia avei et Candidémie; Insuffisance rénale aiguë; Angiocholite

As a result, 323 lines were adjusted by replacing the semicolon with a coma and adding 'NULL' values in the corresponding field, as in:

Example 6.
140646;2014;1;35;6;5;HTA traitée, tabagisme;NULL;NULL
141050;2014;1;80;2;5;état grabataire, démence mixte , DNID;NULL;NULL

Detecting Multiple Causes in Statements As stated previously, there are cases of multiple code assignment (2.2) whose detection may improve the recall score. We didn't find in last year's experiments any explicit definition of a strategy dedicated to multi-code classification even though 27% of the statements in the training sets were multi-coded. Examples of multiple causes in one statement are given below in 7.

Example 7.

Statement	**#causes codes**
Bronchopneumopathies d'inhalation à répétition et apraxie de la déglutition	2 690,R13
Alcoolisme chronique (stéatose hépatique, cardiomyopathie dilatée)	3 F102,I420,K760
Carbonisation diffuse avec traumatisme thoracique	2 T293,S299

On the basis of a manual analysis of repeated lines in the training set, a list of 17 potential causes separators were identified (Table 3). All statements containing at least one of the potential separators were systematically analyzed in order to keep the 7 most reliable separators for our splitting method. The statement is segmented each time a reliable separator is detected.

Table 3. Potential cause separators—in color those actually used by the system

Separator type	**tokens**			
punctuation	,	.	/	+
simple coordinator	et/ou	et	ou	
simple preposition	chez	par	sur	avec
time preposition	puis	après		
time adverbial	au stade d'	au cours d'	suivi de	suite à

14,867 raw statements were segmented by the splitter and 14,857 aligned ones[9]. The performance of the splitter has not been evaluated on the training data but a comparison with the gold standard aligned data indicates that more than 88% of the segmented aligned statements are repeated at least once in the gold standard (13,173/14,857).

[9] We henceforth simply mention 'Raw' and 'Aligned' to refer to the raw test set vs. the aligned test set.

Spell Checking and Normalization A spellchecker was used to handle four types of typing variations: diacritics, punctuation, compound word separators and some of the most frequent typing errors.

Words with diacritics are very often misspelled in French (as illustrated in ex 1). We adopted a radical strategy for solving such misspelling errors consisting of replacing all diacritics characters with their non-diacritic forms. Because punctuations were not considered relevant for the task we decided to remove all punctuation marks from the raw texts once multiple causes have been detected.

As for compound words normalization, we applied the method developed by the LIMSI team last year [17]. It consists in:

1. extracting from the training and development sets all the hyphenated words (e.g. *cardio-respiratoire*),
2. generating their concatenated counterpart (e.g. *cardiorespiratoire*),
3. generating their split counterpart (e.g. *cardio respiratoire*),
4. building a lexicon in which hyphenated and concatenated compound words are aligned with the split version (*cardiorespiratoire* $\Rightarrow$ *cardio respiratoire*),
5. applying this lexicon on test data by replacing all variants with their split form.

The resulting lexicon contains 2,155 entries. The normalization method processed almost 10% of the lines (9,188 raw lines and 9,169 aligned lines); 9,606 raw and 9,585 aligned tokens were replaced.

During the building of the compound words lexicon we observed some recurrent typing errors on biomedical entities, such as *ardio-respiratoire* instead of *cardio respiratoire*. This analysis was completed by the systematic observation of all the raw texts associated with the five most frequent ICD10 codes:

- 6260 rawTexts coded R092 (*Arrêt cardio respiratoire* i.e. 'cardiorespiratory failure'),
- 4576 rawTexts coded A419 ('Septicaemia'),
- 4606 rawTexts coded R688 ('Other specified general symptoms and signs'),
- 5075 rawTexts coded I10 ('Essential (primary) hypertension'),
- 4152 rawTexts coded I509 ('Heart failure, unspecified').

The resulting lexicon lists all the observed spelling errors (for example, the 36 spellings of the term 'cardio respiratoire'[10]). It contains 378 entries and was applied to 67 lines in the test set.

[10] ardiorespiratoire, cadiorespiratoire, caerdiorespiratoire, cardiiorespiratoire, cardioeespiratoire, cardioespiratoire, cardiorecpiratoire, cardiorepiratoire, cardiorepsiratoire, cardioresiratoire, cardioresoiratoir, cardioresoiratoire, cardioresperatoire, cardiorespi, cardiorespiaratoire, cardiorespiartoire, cardiorespir, cardiorespira, cardiorespiraoire, cardiorespirapoire, cardiorespiratire, cardiorespiratoie, cardiorespiratoir, cardiorespiratoire, cardiorespiratoitre, cardiorespiratopire, cardiorespiratore, cardiorespiratorie, cardiorespiraztoire, cardiorespireatoire, cardiorespitatoire, cardiorespitoire, cardiorespitratoire, cardiorespratoire, cardiorespriratoire, cardirespiratoire, cardoirespiratoire, cardorespiratoire, cerdiorespiratoire.

4.2 Query Expansion

Before querying the indexed collection with the preprocessed lines, we opted for a two step query expansion procedure.

Abbreviations and Acronyms The first method consists in expanding all abbreviations and acronyms occurring in the raw texts with their full form. A first list of 60 single word abbreviations and acronyms was extracted from the analysis of the 1,000 most frequent words (covering almost 90% of the total words) and associated to their full form. 22 additional abbreviations came up from the observation of the raw texts associated with the five most frequent ICD10 codes described above, such as:

- anat path ⇒ anatomo pathologie
- post op ⇒ post operatoire
- cardio respi ⇒ cardio respiratoire

The resulting lexicon contains 82 entries. This expansion procedure processed slightly more than 8% lines (7,506 raw lines and 7,435 aligned lines); 9,335 raw and 9,210 aligned tokens were expanded. Table 4 indicates the scores with and without this expansion, evaluated on a sample of 200 statements extracted from the development set. It shows that all scores (precision and recall) increase by 3 points when abbreviations and acronyms expansion is applied.

Table 4. Abbreviations and acronyms processing: scores on 200 statements extracted from the development data set

	Precision	Recall	F-measure
Without disabbreviation	0.6400	0.4604	0.5356
With disabbreviation	**0.6750**	**0.4856**	**0.5649**

Expanding with Similar and Associated Terms The second expansion technique aims at adding words considered similar or associated to the words occurring in the raw texts. A handmade lexicon was built on the basis of the analysis of the 1,000 most frequent words as described in the previous section. 4 types of relations were considered:

- morphological relations between adjectives and their related lexical root: *cardio* ⇒ *coeur* (*heart*), *cardiaque* ⇒ *coeur*, *vasculo* ⇒ *vaisseaux* (*vessel*), etc.;
- morphological relations between derived nouns and their related lexical root: *alcoolisme* ⇒ *alcool*;

- hypernym relations between the term *opération* (*procedure*) and all its potential of manually detected hyponyms: *opération* ⇒ *amputation*, *colectomie*, *pancréatectomie*;
- meronym relations between the term 'cancer' and some of its symptoms or associated pathologies: *cancer* ⇒ *métastase*, *tumeur*, *carcinome*;
- synonym relations between various observed terms: *nombril* (*navel*) ⇒ *ombilic* (*ombilicus*), *partum* ⇒ *accouchement* (*Childbirth*), *post* ⇒ *après* (*after*), etc.

The resulting lexicon contains 285 entries. This expansion procedure matched almost 31% of the lines (28,428 raw lines and 28,346 aligned lines); 34,453 raw and 34,326 aligned tokens were expanded. Table 5 indicates the scores with or without this expansion, evaluated on a sample of 200 statements extracted from the development set. Even if this technique does not improve the scores, we still decided to apply it to test set.

Table 5. Syn processing: scores on 200 statements extracted from the development data set

	Precision	Recall	F-measure
Without synonyms	0.6750	0.4856	0.5649
With synonyms	0.6750	0.4856	0.5649

5 Results and Discussion

In this last section we present the results obtained by our system on the task's test data. The first code provided by Solr has been considered, even in case of equally ranked codes. Table 6 indicates that our F1 scores are above the average (of submitted runs) for raw data coding and slightly under it for aligned data coding. Run1, with split multiple causes, gets higher scores, in particular the recall, only for the aligned data set.

The system presented here is the result of a student project. The students enthusiastically participated in the competition and the teachers took the opportunity of this context to bring them to work in group, explore various NLP techniques and observe data. A large amount of data was analyzed, mostly qualitatively, with basic corpus linguistics strategies (observation of the 1000 most frequent words and of the raw texts of the 5th most frequent ICD10 codes). Many improvements would be needed to go beyond this first step: a number of ideas have been explored but would need in-depth study to be generalized. The collection building should be reconsidered, especially by taking into account all the ICD10 codes and by pursuing the idea of a distinction between the lines which express the direct causes and the lines relative to indirect causes. Solr parameters would need to be more thoroughly exploited and evaluated. Many

Table 6. LITL team scores on test data

FR raw-ALL	**Precision**	**Recall**	**F-measure**
LITL-run2	**0.67**	0.41	0.51
LITL-run1	0.65	0.40	0.50
average	0.47	0.36	0.41
median	0.54	0.41	0.51
FR aligned-ALL	**Precision**	**Recall**	**F-measure**
LITL-run1	0.61	**0.55**	**0.58**
LITL-run2	**0.65**	0.40	0.50
average	0.65	0.56	0.59
median	0.63	0.54	0.55

proposals were made for preprocessing stages and some of them should be deepened and extended: spellchecking, normalization, cause splitting have been only partially handled, focusing on some parts of the vocabulary, without sufficient evaluation.

Acknowledgements

We would like to thank Ludovic Tanguy for his contribution to the supervision of this student project.

References

1. Blanquet, A., Zweigenbaum, P.: A lexical method for assisted extraction and coding of icd-10 diagnoses from free text patient discharge summaries. In: Proceedings of the AMIA Symposium. p. 1029. American Medical Informatics Association (1999)
2. Cabot, C., Soualmia, L.F., Dahamna, B., Darmoni, S.J.: Sibm at clef ehealth evaluation lab 2016: Extracting concepts in french medical texts with ecmt and cimind. In: Proceedings of CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (2016)
3. Dermouche, M., Looten, V., Flicoteaux, R., Chevret, S., Velcin, J., Taright, N.: Ecstra-inserm@ clef ehealth2016-task 2: Icd10 code extraction from death certificates. In: Proceedings of CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (2016)
4. Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: Clef 2017 ehealth evaluation lab overview. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings. vol. 10439. springer (2017)
5. Ho-Dac, L.M., Tanguy, L., Grauby, C., Mby, A.H., Malosse, J., Rivière, L., Veltz-Mauclair, A., Wauquier, M.: Litl at clef ehealth2016: recognizing entities in french biomedical documents. In: Proceedings of CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (2016)

6. Lavergne, T., Névéol, A., Robert, A., Grouin, C., Rey, G., Zweigenbaum, P.: A dataset for icd-10 coding of death certificates: Creation and usage. In: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2016). pp. 60–69. Osaka, Japan (2016)
7. Mottin, L., Gobeill, J., Mottaz, A., Pasche, E., Gaudinat, A., Ruch, P.: Bitem at clef ehealth evaluation lab 2016 task 2: Multilingual information extraction. In: Proceedings of CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (2016)
8. Mujtaba, G., Shuib, L., Raj, R.G., Rajandram, R., Shaikh, K., Al-Garadi, M.A.: Automatic icd-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. PLOS ONE 12(2), 1–27 (02 2017), https://doi.org/10.1371/journal.pone.0170242
9. Névéol, A., Anderson, R.N., Bretonnel Cohen, K., Cohen, K., Grouin, C., Lavergne, T., Rey, G., Robert, A., Rondet, C., Zweigenbaum, P.: Clef ehealth 2017 multilingual information extraction task overview: Icd10 coding of death certificates in english and french. In: CLEF 2017 Labs Working Notes. CEUR-WS (2017)
10. Névéol, A., Goeuriot, L., Kelly, L., Cohen, K., Grouin, C., Hamon, T., Lavergne, T., Rey, G., Robert, A., Tannier, X., Zweigenbaum, P.: Clinical information extraction at the clef ehealth evaluation lab 2016. In: Proceedings of CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (2016)
11. Pavillon, G., Laurent, F.: Certification et codification des causes médicales de décès. Bulletin épidémiologique hebdomadaire 30(31), 134–138 (2003)
12. Shah, A.D., Martinez, C., Hemingway, H.: The freetext matching algorithm: a computer program to extract diagnoses and causes of death from unstructured text in electronic health records. BMC Medical Informatics and Decision Making 12(1), 88 (2012)
13. Tanguy, L., Tulechki, N., Urieli, A., Hermann, E., Raynal, C.: Natural language processing for aviation safety reports: from classification to interactive analysis. Computers in Industry 78, 80–95 (2016)
14. Van Mulligen, E., Afzal, Z., Akhondi, S.A., Vo, D., Kors, J.A.: Erasmus mc at clef ehealth 2016: Concept recognition and coding in french texts. In: Proceedings of CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (2016)
15. Wang, A.Y., Sable, J.H., Spackman, K.A.: The SNOMED clinical terms development process: refinement and analysis of content. In: Proceedings of the AMIA Symposium. p. 845. American Medical Informatics Association (2002)
16. Wingert, F., Rothwell, D., Côté, R.A.: Computerised Natural Medical Language Processing for Knowledge Engineering, chap. Automated indexing into SNOMED and ICD, pp. 201–239. North-Holland, Amsterdam (1989)
17. Zweigenbaum, P., Lavergne, T.: Limsi icd10 coding experiments on cépidc death certificate statements. In: Proceedings of CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (2016)