SIBM at CLEF eHealth Evaluation Lab 2017: Multilingual Information Extraction with CIM-IND

Chloé Cabot[1], Lina F. Soualmia[1,2], and Stéfan J. Darmoni[1,2]

[1] Normandie Univ., SIBM, TIBS - LITIS EA 4108, Rouen University and Hospital, France
[2] French National Institute for Health, INSERM, LIMICS UMR-1142, France
chloe.cabot@chu-rouen.fr, lina.soualmia@chu-rouen.fr, stefan.darmoni@chu-rouen.fr

Abstract. This paper presents SIBM's participation in the Task 1: Multilingual Information Extraction - ICD10 coding of the CLEF eHealth 2017 evaluation initiative which focuses on named entity recognition in French and English death certificates. We addressed the identification of relevant clinical entities within the International Classification of Diseases version 10 (ICD10) in the CépiDC and CDC datasets with our CIM-IND system. CIM-IND is a multilingual system designed to recognize named entities in French and English texts using a dictionary-based approach and natural language processing and fuzzy matching methods. The evaluation was performed for two cases: (i) for all ICD10 codes, the main evaluation for the task and (ii) for ICD10 codes addressing a particular type of deaths, called external causes or violent deaths. On the English test set, our system obtained F-scores of 0.81 for all ICD10 codes and 0.4066 for external causes. On the French aligned test set, our system obtained F-scores of 0.8038 for all ICD10 codes and 0.5011 for external causes. On the French raw test set, our system obtained F-scores of 0.7636 for all ICD10 codes and 0.4897 for external causes. These scores were substantially higher than the average score of the systems that participated in the challenge.

Keywords: Information extraction; Entity recognition; Lexical semantics; Natural Language Processing; International Classification of Diseases

1 Introduction

Since the amount of digital medical documents has widely expanded in the last twenty years, the information retrieval from such heterogeneous documents has become a significant challenge to address a large variety of tasks in clinical and biomedical research as well as personalized medicine. Named entity recognition (NER) is a basic sub-task of information extraction that aims to extract and classify entity names from text. The NER problem has been studied widely in

the last decade in the biomedical field as well as others such as social media [1] or speech data [2]. As the use of NER services has expanded, state-of-the-art algorithms have improved on formal medical text for English [3]. However, NER algorithms struggle to adapt to free text because algorithms are designed for formal text and are based on features present in well-formed text such as biomedical articles. Free text in medical notes comprises spelling errors, incorrect use of punctuation, grammar and capitalization [4]. In other languages, free text can also present incorrect use of diacritical marks. In medical reports, text is usually made from short or incomplete sentences, similar to note-taking, with a substantial use of ambiguous abbreviations. Usually, clinical records are created in a rush without any proofing. Consequently, a large number of spelling errors occurs. These errors should not only be related to the complexity of the language but also to characteristics of the medical domain. Siklósi et al. found that the most frequent types of errors are the unintentional mistyping, grammatical errors, sentence fragments, and non-standardized abbreviations [5]. In fact, as opposed to formal text, abbreviations are rarely defined in medical reports. Despite the efforts made in NER, even in the biomedical domain, information extraction in clinical notes still has to undertake several challenges [6].

Since 1995, the department of BioMedical Informatics of the Rouen University Hospital (SIBM, URL: www.cismef.org) has been working on developing tools to access health knowledge (information retrieval and automatic indexing) in French [7–10]. More recently, our team has worked on the evaluation of health information systems and information retrieval and indexing in Electronic Health Records (EHRs) [11, 12]. In this context, a multilingual system called CIM-IND has been developed. CIM-IND is designed to recognize named entities in French and English texts using a dictionary-based approach and natural language processing and fuzzy matching methods. The main objective of this system is to deal accurately and efficiently with the informal and noisy nature of free text in medical reports. To assess the performance of CIM-IND, our team participated in the CLEF eHealth 2016 Task 2 [13, 14] which aimed at fully automatically identify clinically relevant entities in death certificates in French and obtained average results [15]. While death certificates are standardized documents filled by physicians to report the death of a patient, they usually present spelling or typing errors, abbreviations, and, in French, non-diacritized text or a mix of cases and diacritized text. The main motivation in participating is to improve the functionalities of the tool and to determine the progress achieved since our last year participation and our ability to address the issues detected then. As the Task 1: Multilingual Information Extraction - ICD10 coding of the CLEF eHealth 2017 evaluation initiative involved assigning codes from the International Classification of Diseases, version 10 (ICD10) to both French and English death certificates[16, 17], we were also able to test our multilingual approach.

The rest of the paper is organized as follows. In Section 2 we introduce our extraction approach and tools used in this task and we describe our experimental setup. Section 3 reports on our results. Section 4 presents some error analyses and reflections and wraps up concluding remarks and outlines future work.

2 Material and methods

2.1 Test datasets

French CépiDC datasets Since 1968, the CépiDC, a French National Institute for Health and Medical Research (Inserm) laboratory, is dedicated to elaborate annually the national medical causes of death statistics in association with the French National Institute for Statistics and Economic Studies (Insee), the dissemination of the data and the studies and researches on the medical causes of death. These statistics are built from information from death certificates. The CépiDC team handles a database containing more than 18,000,000 death records [18]. The task consists of extracting ICD10 codes from the raw lines of death certificate text. The task is an information extraction task that relies on the text supplied to extract ICD10 codes from the certificates, line by line. Two datasets are provided for the task. The first dataset is called "aligned dataset" and the second is called "raw dataset". As the structure of the files provided by these two sets differs, some minor adjustments were necessary to process them.

Aligned dataset The dataset includes 31,690 death certificates processed by CépiDC in 2014 totalling 91,962 lines. The annotations in the CépiDC corpus consist of ICD10 codes and were assigned per text line.. The dataset is supplied in one CSV-formatted file. Each row contains twelve information fields associated with a raw line of text from an original death certificate as follows:

- DocID: death certificate ID
- YearCoded: year the death certificate was processed by CpiDC
- Gender: gender of the deceased
- Age: age at the time of death, rounded to the nearest five-year age group
- LocationOfDeath: Location of death
- LineID: line number within the death certificate
- 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: minutes, hours, days, months, years
- IntValue: length of time the patient had been suffering from coded cause
- CauseRank: Rank of the ICD10 code
- StandardText: dictionary entry or excerpt of the raw text that supports the selection of an ICD10 code (if any)
- ICD10: ICD10 code associated with the certificate corresponding to the DocID and LineID

The output comprises the 9 input fields plus two text fields (CauseRank and StandardText) used to report evidence text supporting the ICD10 code supplied in the twelfth, final field.

Raw dataset The data from 31,683 death certificates is distributed over three CSV-formatted files. The first file includes the following fields: DocID, YearCoded, LineID, RawText, IntType, IntValue. The second files includes the following fields: DocID, YearCoded, Gender, PrimCauseCode, Age, LocationOfDeath. The third file includes the following fields: DocID, YearCoded, LineID.

English CDC dataset The data from 6,665 death certificates is distributed over three CSV-formatted files. The first file includes the following fields: DocID, YearCoded, LineID, RawText, IntType, IntValue. The second file includes the following fields: DocID, YearCoded, Gender, PrimCauseCode, Age, LocationOfDeath. The third file includes the following fields: DocID, YearCoded, LineID.

2.2 Dictionaries

The French CépiDC corpus includes six versions of a manually curated ICD10 dictionary developed at CépiDC corresponding to years: 2006-2010, 2011, 2012, 2013, 2014 and 2015. The English CDC corpus includes a manually curated ICD10 dictionary developed by the CDC providing 170,285 entries. These resources were used to build spelling dictionaries. Moreover, the training sets were used to complete these dictionaries.

Spelling dictionaries For each language, the dictionary versions were merged if necessary. Each ICD term was split into words and duplicates removed. The two lists of unique words obtained provided a spelling dictionary for each language.

Additional dictionaries Then, an additional dictionary was computed from each training set by extracting ICD10 code and term combinations. The number of times an ICD10 code was used in the training corpus was also determined. For ambiguous terms, i.e. terms that corresponded with more than one ICD10 code, the most used term was kept. Each additional dictionary was merged with dictionaries provided in the corresponding corpus. If a term was present in both the additional dictionary and a corpus dictionary but the corresponding codes were different, the code from the additional dictionary was removed to avoid introducing ambiguity between dictionary versions. This processing helped to complete the provided dictionaries especially with some lacking abbreviations.

2.3 Extracting ICD10 concepts from death certificates with CIM-IND

CIM-IND is designed to match ICD10 terms from the text as input in the relevant version of the ICD10. The extraction is performed at the phrase level of the text using natural language processing techniques. The system is built using Python and Python/C extensions and provides a response in CSV format for each identified concept with: (i) the entry text, (ii) the offset of the first and the

final word contained in the health concept, (iii) the ICD10 identifier and (iv) the ICD10 term. CIM-IND performs three main steps to identify ICD10 terms: normalization, candidate selection and candidate ranking.

Normalization Several pre-processing steps are performed, including stop words filtering (using the default NLTK stop word lists for both French and English [19]) and elision filtering (removing abbreviated articles that are contracted with terms). Words are matched case-insensitive. Diacritics in French texts are conserved and Unicode is used for matching. Finally, spell checking is performed with the Enchant library using the manually built dictionary.

Candidate selection A method based on the phonetic encoding algorithm Double Metaphone (DM) [20] is used to operate a first approximate term search. The DM phonetic encoding algorithm is the second generation of the Metaphone algorithm. It is designed primarily to encode American English names while taking into account the fact that such words can have more than one acceptable pronunciation. Double Metaphone can compute a primary and a secondary encoding for a given word or name to indicate both the most likely pronunciation as well as an optional alternative pronunciation (hence the "double" in the name). DM tries to account for myriad irregularities in English as well as Slavic, Germanic, Celtic, Greek, French, Italian, Spanish, Chinese, and other languages. Though powerful, DM does have its limitations and drawbacks. DM was designed for searching lists of proper names rather than large amounts of text. DM may not match grossly misspelled words that seriously alter the phonetic structure of the word. Despite its limitations, the DM algorithm, which is free to use and open source, still holds as a flexible and powerful phonetic encoding system today, especially in a multilingual approach.

First, CIM-IND computes DM encoding for each word included in the normalized phrase. Then, ICD10 term candidates with matching DM encoding are retrieved. This step provides quickly a list of relevant ICD10 term candidates and allows to perform time-consuming analyses on a reduced set of terms in the final step. In this way, our system relies on a database to store pre-computed DM encoding for each word available in each ICD10 version dictionary.

Candidate ranking Finally, a Weighted Distance Score (WDS) algorithm has been developed to rank the list of candidate terms. The WDS algorithm returns a similarity score scaled from 0 to 100 for each candidate, 100 representing a perfect match. The most likely term having the highest score is retained as the matching ICD10 term. As only one or multiple ICD10 terms can be present in a phrase, two cases are considered. First, if the candidate sequence s_1 length is similar to the processed line s_2 length (i.e only one ICD10 term is expected), two scores are computed: (i) a base score (BS) and (ii) a set score (SeS). The BS is computed by determining the Levenshtein distance between the sequences s_1 and s_2 scaled from 0 to 100. The SeS finds all alphanumeric tokens in each

string and treats them as a set. Then two strings are constructed by concatenate, on the one hand, the sorted intersection and, on the other hand, the sorted remainder. Then, the distance of these strings are computed controlling any unordered partial matches.

Else, if one of the sequences is 1.5 times longer than the other, two partial scores are computed: (i) a partial base score (PBS) and (ii) a partial set score (PSeS). The PBS returns the distance of the most similar substring as a number between 0 and 100. First each block representing a sequence of matching characters in a string is determined. Then, the best partial match will be the one aligning with at least one of those blocks. The PSeS computes the PBS for each string built from the sorted intersection and the sorted remainder of s_1 and s_2. To assure that only full results can return a perfect match, partial scores are scaled based on the length of s_1 and s_2. All set scores are scaled by 0.95. Finally, the WDS score is determined as the highest of these scores.

```
64185;2013;2;85;2;5;SYNDROME DE GLISEMENT AVEC GRABATISATION DEPUIS
OCTOBRE 2012;4;3;6-1;syndrome glissement;R453
64185;2013;2;85;2;5;SYNDROME DE GLISEMENT AVEC GRABATISATION DEPUIS
OCTOBRE 2012;4;3;6-1;grabatisation;R263
79317;2013;2;85;2;6;héùorragie digestive basse sur surdosage en
AVK;3;5;6-3;hémorragie digestive basse;K921
79317;2013;2;85;2;6;héùorragie digestive basse sur surdosage en
AVK;3;5;6-3;surdosage avk;X44
64370;2013;1;80;2;5;ABCES CERVICAL . LARYNGECTOMIE TOTALE.ATCD
D'IDM.;NULL;NULL;;laryngectomie totale;Z900
64370;2013;1;80;2;5;ABCES CERVICAL . LARYNGECTOMIE TOTALE.ATCD
D'IDM.;NULL;NULL;;abcès cervical;L021
64370;2013;1;80;2;5;ABCES CERVICAL . LARYNGECTOMIE TOTALE.ATCD
D'IDM.;NULL;NULL;;antécédent infarctus myocarde;I258
```

Fig. 1. Annotation file in CSV containing ICD10 concepts extracted with CIM-IND in French

Figure 1 gives an example of processing French texts with CIM-IND. The seventh field contains the text to annotate, the eleventh the ICD10 dictionary entry matching the text and the last field the corresponding ICD10 code. Similarly, Figure 2 gives an example of processing English texts with CIM-IND.

For example, in Figure 1, lines 1-2 contains the misspelled word "glisement" (for French "glissement") and lines 3-4 contains the misspelled word "héúorragie" (for French "hémorragie"). This first error is correctly processed by the DM algorithm providing the same encoding for both the misspelled and correct words. However, the second error is not properly processed. As the misspelling profoundly alters the phonetic of the word, the DM algorithm processes a different encoding than for the correct word. This highlights the importance to process

```
13496;2015;;;;6;Senile dementia of Alzheimer's type ASHD;;;;senile
dementia;F03
13496;2015;;;;6;Senile dementia of Alzheimer's type
ASHD;;;;alzheimer;G309
13496;2015;;;;6;Senile dementia of Alzheimer's type ASHD;;;;ashd;I251
16915;2015;;;;2;HEALTHCAREASSOCIATED PNEUMONIA;;;;healthcare-associated
pneumonia;J189
```

Fig. 2. Annotation file in CSV containing ICD10 concepts extracted with CIM-IND in English

a spell checking of the normalized text to avoid grossly misspelled words before the DM processing and so secure a proper list of candidates.

Regarding execution time, CIM-IND is able to process a line from 50 to 300 ms depending on its length.

3 Results

3.1 French CépiDC datasets

CIM-IND was run on both French test sets and one run was submitted for each of these datasets. Table 1 shows the results obtained on the raw dataset together with the average and median performance scores of the runs of all task participants. Table 2 shows the results obtained on the aligned dataset.

On the raw dataset, CIM-IND achieved a precision of 0.8568 and a recall of 0.6886 ($F1 = 0.7636$) for all ICD10 codes. Regarding only ICD10 codes corresponding to external causes (meaning violent deaths), CIM-IND achieved a substantial lower performance with a precision of 0.567 and a recall of 0.431 ($F1 = 0.4897$).

On the aligned dataset, CIM-IND achieved a precision of 0.8346 and a recall of 0.7751 ($F1 = 0.8038$) for all ICD10 codes. Regarding only ICD10 codes corresponding to external causes, CIM-IND achieved again a lower performance with a precision of 0.5343 and a recall of 0.4717 ($F1 = 0.5011$).

Since the main difference between these two datasets was related to formatting, it was expected to obtain quite similar results. However, remarkably, the aligned dataset obtains a higher recall than the raw dataset. Then, it should be noted that performance is considerably lower regarding only external causes related ICD10 codes for both test sets. Overall, our performance results are considerably better than the average and median score of all submitted runs.

3.2 English CDC dataset

One run was submitted for the English CDC set. Table 3 shows the results obtained on this dataset together with the average and median performance scores of the runs of all task participants.

CIM-IND achieved a precision of 0.8393 and a recall of 0.7827 ($F1 = 0.81$) for all ICD10 codes. Regarding only ICD10 codes corresponding to external causes, CIM-IND achieved a lower performance with a precision of 0.4261 and a recall of 0.3889 ($F1 = 0.4066$).

Regarding all ICD10 codes, these results are slightly better than the results obtained with the French raw dataset but remarkably similar to those obtained with the aligned dataset. Again, there is a significant performance drop regarding only external causes related ICD10 codes. In this case, results are lower than those obtained on both French datasets, for both precision and recall. Overall, in both evaluations, our results are higher than the average and median score of all submitted runs.

Table 1. ICD10 coding performance on the French CépiDC raw test dataset

	All causes			External causes		
	Precision	Recall	F-measure	Precision	Recall	F-measure
SIBM-run1	0.8568	0.6886	0.7636	0.5670	0.4310	0.4897
average	0.4747	0.3583	0.4059	0.3668	0.2474	0.2921
median	0.5411	0.4136	0.5080	0.4431	0.2834	0.3764

Table 2. ICD10 coding performance on the French CépiDC aligned test dataset

	All causes			External causes		
	Precision	Recall	F-measure	Precision	Recall	F-measure
SIBM-run1	0.8346	0.7751	0.8038	0.5343	0.4717	0.5011
average	0.6479	0.5555	0.5933	0.5051	0.3109	0.3663
median	0.6288	0.5396	0.5484	0.5080	0.3330	0.4056

Table 3. ICD10 coding performance on the English CDC test dataset

	All causes			External causes		
	Precision	Recall	F-measure	Precision	Recall	F-measure
SIBM-run1	0.8393	0.7827	0.8100	0.4261	0.3889	0.4066
average	0.6549	0.5586	0.6017	0.3986	0.2749	0.2549
median	0.6459	0.5267	0.5892	0.2791	0.2619	0.2740

4 Discussion and conclusion

The development of CIM-IND started last year and the system was evaluated in the corresponding CLEF eHealth 2016 task, only on one French corpus. In 2016, CIM-IND obtained a F1 score of 0.6795, which was slightly below the average results [15]. Since then, various improvements have been developed concerning especially the ranking of ICD10 term candidates and CIM-IND's ability to deal with free text inconsistencies. This year's results have demonstrated these improvements with a 12% increase in F1 score in the French raw dataset and an 18% increase in F1 score in the French aligned dataset. Moreover, this year's challenge demonstrated that CIM-IND performed broadly as well in both English and French, achieving above-average results in both languages.

However, some aspects of our results should be investigated. Although CIM-IND achieved satisfactory results, we noticed that some errors due to disambiguation or misspellings and inconsistencies remain. In particular, significant misspellings occurring on words which are not part of the spelling dictionary would result in incorrect DM encoding, and so an improper list of candidate terms.

In English text, our results could be slightly improved with a more complete terminology or a larger training set to cover some missing terms, especially abbreviations. Moreover, the performance drop regarding external causes-related ICD10 codes should be investigated and seems to affect all submitted runs. External causes present a specific context and often a specific terminology related to accidents, violent deaths or treatment-induced overdoses. They occur more rarely in the training sets. Actually only 2440 lines in the French training set (110,869 lines) and 313 lines in the English train set (39,333 lines) appear to be related to external causes (ICD10 codes V01 to Y98). This can explain the reduced performance to some extent. Also, in some cases, the ICD10 codes associated with a given line use the context provided in other lines of the same death certificate. CIM-IND processes each line independently and then was not able to properly annotate such lines.

The main conclusion of this work and the obtained results is that improvements can still be performed to enhance first the processing of the given terminologies and disambiguation-related issues and also the recognition and processing of spelling errors. We plan on deepening these two aspects and to participate to other challenges in the future to keep track of our developments.

References

1. Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., Bontcheva, K.: Analysis of named entity recognition and linking for tweets. Information Processing & Management **51**(2) (March 2015) 32–49
2. Ma, Y., Kim, J.j., Bigot, B., Khan, T.M.: Feature-enriched word embeddings for named entity recognition in open-domain conversations. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2016) 6055–6059

3. Mork, J., Aronson, A., Demner Fushman, D.: 12 years on - Is the NLM medical text indexer still useful and relevant? Journal of biomedical semantics **8**(1) (February 2017) 8
4. Lai, K.H., Topaz, M., Goss, F.R., Zhou, L.: Automated misspelling detection and correction in clinical free-text records. Journal of biomedical informatics **55** (June 2015) 188–195
5. Siklósi, B., Novák, A., Prószéky, G.: Context-aware correction of spelling errors in Hungarian medical documents. Computer Speech & Language **35** (January 2016) 219–233
6. Menasalvas, E., Gonzalo-Martin, C.: Challenges of Medical Text and Image Processing: Machine Learning Approaches. In: Machine Learning for Health Informatics. Springer International Publishing, Cham (2016) 221–242
7. Darmoni, S.J., Thirion, B., Leroyt, J.P., Douyère, M., Lacoste, B., Godard, C., Rigolle, I., Brisou, M., Videau, S., Goupyt, E., Piott, J., Quéré, M., Ouazir, S., Abdulrab, H.: A search tool based on 'encapsulated' MeSH thesaurus to retrieve quality health resources on the internet. Medical informatics and the Internet in medicine **26**(3) (July 2001) 165–178
8. Neveol, A., Rogozan, A., Darmoni, S.: Automatic indexing of online health resources for a French quality controlled gateway. Information Processing & Management **42**(3) (May 2006) 695–709
9. Soualmia, L.F., Sakji, S., Letord, C., Rollin, L., Massari, P., Darmoni, S.J.: Improving information retrieval with multiple health terminologies in a quality-controlled gateway. Health Information Science and Systems **1**(1) (2013) 8
10. Chebil, W., Soualmia, L.F., Omri, M.N., Darmoni, S.J.: Indexing biomedical documents with a possibilistic network. JASIST **67**(4) (2016) 928–941
11. Cabot, C., Lelong, R., Grosjean, J., Soualmia, L.F., Darmoni, S.J.: Retrieving Clinical and Omic Data from Electronic Health Records. Stud Health Technol Inform **221** (2016) 115
12. Lelong, R., Cabot, C., Soualmia, L.F.: Semantic Search Engine to Query into Electronic Health Records with a Multiple-Layer Query Language. In: Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR). (2016)
13. Kelly, L., Goeuriot, L., Suominen, H., Neveol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2016. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Springer, Cham (September 2016) 255–266
14. Névéol, A., Goeuriot, L., Kelly, L.: Clinical information extraction at the CLEF eHealth evaluation lab 2016. In: Proceedings of CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. (2016)
15. 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: CEUR-WS Working Notes of the Conference and Labs of the Evaluation Forum CLEF. (2016) 47–60
16. Goeuriot, L., Kelly, L., Suominen, H., Neveol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: CLEF 2017 eHealth Evaluation Lab Overview. In: CLEF - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science LNCS, Springer. (September 2017)
17. Neveol, A., Anderson, R.N., Cohen, K.B., 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 Evaluation Labs and Workshop Online Working Notes, CEUR-WS. (September 2017)

18. Pavillon, G., Laurent, F.: Certification et codification des causes médicales de décès. Bulletin épidémiologique hebdomadaire (2003)
19. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media, Inc (2009)
20. Philips, L.: The double metaphone search algorithm. C/C++ Users Journal **18**(6) (June 2000) 38–43