Erasmus MC at CLEF eHealth 2016: Concept Recognition and Coding in French Texts

Erik M. van Mulligen, Zubair Afzal, Saber A. Akhondi, Dang Vo, and Jan A. Kors

Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
{e.vanmulligen,m.afzal,s.ahmadakhondi,v.dang,j.kors}
@erasmusmc.nl

Abstract. We participated in task 2 of the CLEF eHealth 2016 challenge. Two subtasks were addressed: entity recognition and normalization in a corpus of French drug labels and Medline titles, and ICD-10 coding of French death certificates. For both subtasks we used a dictionary-based approach. For entity recognition and normalization, we used Peregrine, our open-source indexing engine, with a dictionary based on French terms in the Unified Medical Language System (UMLS) supplemented with English UMLS terms that were translated into French with automatic translators. For ICD-10 coding, we used the Solr text tagger, together with one of two ICD-10 terminologies derived from the task training material. To reduce the number of false-positive detections, we implemented several post-processing steps. On the challenge test set, our best system obtained F-scores of 0.702 and 0.651 for entity recognition in the drug labels and in the Medline titles, respectively. For entity normalization, F-scores were 0.529 and 0.474. On the test set for ICD-10 coding, our system achieved an F-score of 0.848 (precision 0.886, recall 0.813). These scores were substantially higher than the average score of the systems that participated in the challenge.

Keywords: Entity recognition, Concept identification, ICD-10 Coding, Term translation, French terminology

1 Introduction

The CLEF eHealth Evaluation Lab 2016 task 2 consisted of two main subtasks [1, 2]: recognition and normalization of entities in a French biomedical corpus, and coding of French death certificates. The entity normalization had to be based on the Unified Medical Language System (UMLS) [3], and involved assigning UMLS concept unique identifiers (CUIs) to the recognized entities. This task was also used in CLEF eHealth 2015 [4]. The coding task was new and involved the assignment of codes from the International Classification of Diseases, version 10 (ICD-10) [5] to the death certificates. Both tasks had to be performed fully automatically.

We addressed these tasks using dictionary-based indexing approaches. For entity recognition and normalization we used the system that we developed for the same task in the CLEF eHealth 2015 challenge [6], but we trained it on the data that was made available in this year's challenge. Central in our approach is indexing with French terminologies from the UMLS supplemented with automatically translated English UMLS terms, followed by several post-processing steps to reduce the number of false-positive detections. For ICD-10 coding, we used a terminology that was constructed based on the training data and again applied post-processing to improve precision. We describe our systems and their evaluation for each subtask. On the test data, our results for both tasks are well above the average performance of the systems that participated in the CLEF eHealth 2016 task 2 challenge [2].

2 Methods

In the following, we describe the corpora, terminologies, indexing, and post-processing steps separately for each subtask..

2.1 Corpora

Entity recognition and normalization. The training and test data are based on the Quaero medical corpus, a French annotated resource for medical entity recognition and normalization [7]. The Quaero corpus consists of three sub-corpora: titles from French Medline abstracts, drug labels from the European Medicines Agency (EMEA), and patents from the European Patent Office. For the CLEF eHealth challenge, only Medline titles and EMEA documents were made available. The training set consisted of 1665 Medline titles and 6 full EMEA documents (comprising the training and test data previously released in CLEF eHealth 2015); a new test set contained 834 Medline titles and 4 EMEA documents.

The annotations in the Quaero corpus are based on a subset of the UMLS. An entity in the Quaero corpus was only annotated if the concept belonged to one of the following ten semantic groups (SGs) in the UMLS: Anatomy, Chemicals and drugs, Devices, Disorders, Geographic areas, Living beings, Objects, Phenomena, Physiology, and Procedures. Nested or overlapping entities were all annotated, as were ambiguous entities (i.e., if an entity could refer to more than one concept, all concepts were annotated). Also discontinuous spans of text that refer to a single entity could be annotated.

Coding. The data set for the coding of death certificates is called the CépiDC corpus. Each certificate consists of one or more lines of text and some metadata, including age and gender of the deceased, and location of death. The training set contained 65,843 certificates from the period 2006 to 2012, and the test set contained 27,850 certificates from 2013.

The annotations in the CépiDC corpus consist of codes from the ICD-10 and were assigned per text line. For each code that was assigned by the human coder, a term that supports the selection of the code was provided. This term was an excerpt of the text line or an entry of a coding dictionary (see below). Furthermore, the human coder provided for each code the duration that the deceased had been suffering from the coded cause, and a code rank with respect to the cause of death.

2.2 Terminologies

Entity recognition and normalization. We used the terminology that performed best in the CLEF eHealth 2015 concept recognition task [6]. This terminology was constructed from all French terms in the ten relevant SGs of UMLS version 2014AB (77,995 concepts with 161,910 terms). To increase the coverage of this baseline terminology, English UMLS terms were automatically translated into French. We used two translators, Google Translate (GT) [8] and Microsoft Translator (MT) [9], and only included terms that had the same translation in the baseline terminology. The resultant French terminology contained 136,127 concepts with 386,617 terms. Finally, we expanded the terminology with terms from concepts in the training set that were not recognized by our indexing system (false negatives).

Coding. For coding, we constructed two ICD-10 terminologies. A baseline terminology was made by compiling the terms corresponding with each annotated ICD-10 code in the training corpus. The number of times that each code had been assigned in the training corpus, was also determined. For ambiguous terms, i.e., terms that corresponded with more than one ICD-10 code, the term was removed for those codes that occurred less than half as often as the most frequent code with that term. A second, expanded terminology was based on the baseline terminology, but also incorporated codes and terms from four versions of a manually curated ICD-10 dictionary. These dictionary versions have been developed at the Centre d'épidémiologie sur les causes médicales de décès (CépiDC) [10] and were made available by the task organizers. They contained many additional ICD-10 codes and terms that were not present in the training corpus (and thus were lacking in the baseline terminology). If a term was present in both the baseline terminology and a CépiDC dictionary but the corresponding codes were different, the code in the dictionary version was not included in the expanded terminology to avoid introducing ambiguity. If the term in the baseline terminology was ambiguous (had multiple codes), only the term-code combinations in the baseline terminology that were also present in the CépiDC dictionary were incorporated in the expanded terminology.

2.3 Indexing

Entity recognition and normalization. The Quaero corpus was indexed with Peregrine, our dictionary-based concept recognition system [11]. Peregrine

removes stopwords (we used a small list of (in)definite articles and, for French, partitive articles) and tries to match the longest possible text phrase to a concept. It uses the Lexical Variant Generator tool of the National Library of Medicine to reduce a token to its stem before matching [12]. Peregrine is freely available [13].

Peregrine can find partially overlapping concepts, but it cannot detect nested concepts (it only returns the concept corresponding with the longest term). We therefore implemented an additional indexing step. For each term found by Peregrine and consisting of n words ($n > 1$), all subsets of 1 to n–1 words were generated, under the condition that for subsets consisting of more than one word, the words had to be adjacent in the original term. All word subsets were then also indexed by Peregrine. We did not try to find discontinuous terms since there frequency was very low.

Coding. For the coding task, we employed the open-source Solr text tagger [14], using the ICD-10 terminologies to index the death certificates. Several preprocessing steps were performed, including stopword filtering (using the default Solr stopword list for French), ASCII folding (converting non-ASCII Unicode characters to their ASCII equivalents, if existing), elision filtering (removing abbreviated articles that are contracted with terms), and stemming (using the French Snowball stemmer). Words were matched case-insensitive, except for selected abbreviations that were expanded using a synonym list prior to matching.

2.4 Post-processing

Entity recognition and normalization. To reduce the number of false-positive detections that resulted from the indexing, we applied several post-processing steps. First, we removed terms that were part of an exclusion list. The list was manually created by indexing the French part of the Mantra corpus, a large multilingual biomedical corpus developed as part of the Mantra project [15], and selecting the incorrect terms from the 2,500 top-ranked terms.

Second, for any term-SG-CUI combination and SG-CUI combination that was found by Peregrine in the training data, we computed precision scores: *true positives* / (*true positives* + *false positives*). For a given term, only term-SG-CUI combinations with a precision above a certain threshold value were kept. If multiple combinations qualified, only the two with the highest precision scores were selected. If for a given term none of the found term-SG-CUI combinations had been annotated in the training data, but precision scores were available for the SG-CUI combinations, a term-SG-CUI combination was still kept if the precision of the SG-CUI combination was higher than the threshold. If multiple combinations qualified, the two with the highest precision were kept if they had the same SG; otherwise, only the combination with the highest precision was kept. If none of the SG-CUI combinations had been annotated, a single term-SG-CUI combination was selected, taking into account whether the term was the preferred term for a CUI, and the CUI number (lowest first).

Coding. To reduce the number of false-positive codes that were generated during the indexing step, we computed a precision score for each term-code combination that was recognized by the Solr tagger in the training data: *true positives* / (*true positives* + *false positives*). All codes that resulted from term-code combinations with precision values below a given threshold value, were removed.

3 Results

3.1 Entity recognition and normalization

We indexed the Quaero training data, and added the false-negative terms to our terminology. We then ran the system on the Quaero test data, and submitted two runs for both the entity recognition and normalization tasks: one run using the system with a precision threshold of 0.3 (run1, this threshold was also used in our last-year's submission for the same task [6]), the other with a precision threshold of 0.4 (run2). Table 1 shows our performance results for exact match on the test set.

Table 1. Entity recognition and normalization performance on the Quaero test set

		Entity recognition			Entity normalization		
Corpus	Submission	Precision	Recall	F-score	Precision	Recall	F-score
EMEA	Run1	0.623	0.797	0.699	0.578	0.488	0.529
	Run2	0.634	0.786	0.702	0.588	0.481	0.529
	Average score	0.525	0.411	0.435	0.476	0.322	0.376
	Median score	0.600	0.378	0.444	0.447	0.269	0.315
Medline	Run1	0.617	0.690	0.651	0.562	0.410	0.474
	Run2	0.623	0.678	0.649	0.568	0.404	0.472
	Average score	0.503	0.426	0.446	0.501	0.376	0.429
	Median score	0.617	0.438	0.498	0.493	0.383	0.431

As expected, run1 (the system with the lower precision threshold) has lower precision and higher recall than run2, but the differences are small and the F-scores are nearly identical. The results are well above the average and median of the scores from all runs of the challenge participants.

3.2 Coding

To determine the optimal precision threshold for the coding task, we split the CépiDC training data in two equally-sized sets. Precision scores were generated for all term-code combinations that were found using the expanded ICD-10 terminology in one half of the training data and were then used to filter the recognized term-code combinations in the other half of the training data. Table 2 shows the performance for different threshold values on the second half.

Table 2. ICD-10 coding performance on half of the CépiDC training set for different precision-score thresholds

	ICD-10 coding		
Threshold	Precision	Recall	F-score
0.0	0.732	0.818	0.773
0.1	0.799	0.812	0.806
0.2	0.818	0.808	0.813
0.3	0.844	0.803	0.823
0.4	0.863	0.795	0.827
0.5	0.887	0.770	0.822
0.6	0.893	0.742	0.810
0.7	0.902	0.724	0.803
0.8	0.913	0.704	0.795
0.9	0.932	0.631	0.753

Without precision filtering (threshold 0.0), an F-score of 0.773 (precision 0.732, recall 0.818) was obtained. The highest F-score (0.827) was achieved for a threshold of 0.4, mainly because precision greatly improved (0.863), while recall only slightly deteriorated (0.795). The same optimal threshold value was obtained when we used the baseline ICD-10 terminology.

We submitted two runs on the CépiDC test set, one using the expanded ICD-10 terminology (run1), the other using the baseline terminology (run2). For both runs, precision scores were derived from all the training data and the threshold for precision filtering was set at 0.4. Table 3 shows the performance of our system, together with the average and median performance scores of the runs of all task participants.

Table 3. ICD-10 coding performance on the CépiDC test set

	ICD-10 coding		
Submission	Precision	Recall	F-score
Run1	0.890	0.803	0.844
Run2	0.886	0.813	0.848
Average score	0.788	0.664	0.719
Median score	0.811	0.655	0.700

Our results indicate that the baseline terminology (run2) performed slightly better than the expanded terminology (run1) in terms of F-score. Remarkably, the baseline terminology had higher recall than the expanded terminology. Overall, our performance results, in particular recall, are considerably better than the average and median score of all submitted runs.

4 Discussion

We retrained our system for entity recognition and normalization that we developed for the same task in the CLEF eHealth 2015 challenge, and newly developed a dictionary-based system for ICD-10 coding of death certificates. For both systems, the performance on the test sets proved to be substantially better than the averaged results of all systems that participated in the challenge [2].

Our system for entity recognition and normalization performed better on the EMEA subcorpus than on the Medline subcorpus, primarily because of a higher recall. This is in line with the results for this task in the CLEF eHealth 2015 challenge [4]. However, whereas we had expected the performance of our system to be similar or even better than last year (because of the larger training set this year), results actually were worse, in particular for entity normalization [6]. We are currently investigating what may have caused this performance decrease.

The baseline and expanded ICD-10 terminologies that we developed for our coding system, produced almost similar F-scores. Remarkably, although the expanded terminology contained much more codes and terms than the baseline terminology, recall for the baseline terminology was slightly higher. A probable explanation is that the term disambiguation that we performed when expanding the baseline terminology with the ICD-10 dictionaries supplied by the task organizers, effectively prevented some term-code combinations seen in the training data from being included in the expanded terminology. Moreover, since the codes and terms in the baseline terminology were derived from a very large training set, there may have been few new codes and terms in the test set.

Our coding system achieved a very high precision of 0.886, with a recall of 0.813. Reasons for the lower recall include disambiguation errors, spelling mistakes and typos in the death certificates, and missing terms in the terminology. Also, we noticed that some gold-standard code annotations erroneously corresponded with a line that preceded or followed the line that contained the term to be coded, which resulted in false-negative detections (and possibly false-positive detections that were actually correct). Further improvement of our system may be possible by using better curated terminologies and applying spelling correction techniques.

References

1. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2016. In: CLEF 2016 - 7th Conference and Labs of the Evaluation Forum. Lecture Notes in Computer Science (LNCS). Springer, Heidelberg (2016)
2. 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. CLEF 2016 Online Working Notes, CEUR-WS (2016)
3. Bodenreider, O.: The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic Acids Res. 32, D267–270 (2004)

4. Névéol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., Zweigenbaum, P.: CLEF eHealth Evaluation Lab 2015 Task 1b: Clinical Named Entity Recognition. CLEF 2015 Online Working Notes, CEUR-WS (2015)
5. International Classification of Diseases, http://www.who.int/classifications/icd/en/
6. Afzal, Z., Akhondi, S.A., van Haagen, H.H.B.M., van Mulligen, E.M., Kors, J.A.: Biomedical Concept Recognition in French Text Using Automatic Translation of English Terms. CLEF 2015 Online Working Notes, CEUR-WS (2015)
7. Névéol, A., Grouin, C., Leixa, J., Rosset, S., Zweigenbaum, P.: The QUAERO French Medical Corpus: a Ressource for Medical Entity Recognition and Normalization. In: Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM), pp. 24–30 (2014)
8. Google Translate, https://translate.google.com
9. Microsoft Translator, http://www.bing.com/translator
10. Pavillon, G., Laurent, F.: Certification et Codification des Causes Médicales de Décès. Bulletin Epidémiologique Hebdomadaire. 30/31, 134–138 (2003)
11. Schuemie, M.J., Jelier, R., Kors, J.A.: Peregrine: Lightweight Gene Name Normalization by Dictionary Lookup. Proceedings of the BioCreAtIvE II Workshop; Madrid, Spain. pp. 131–133 (2007)
12. Divita, G., Browne, A.C., Rindflesch, T.C.: Evaluating Lexical Variant Generation to Improve Information Retrieval. Proceedings of the American Medical Informatics Association Symposium, pp. 775–779 (1998)
13. Peregrine Indexer, https://trac.nbic.nl/data-mining
14. Solr Text Tagger, https://github.com/OpenSextant/SolrTextTagger
15. Mantra project website, http://www.mantra-project.eu