Biomedical Concept Recognition in French Text Using Automatic Translation of English Terms

Zubair Afzal, Saber A. Akhondi, Herman H.H.B.M. van Haagen,
Erik M. van Mulligen, and Jan A. Kors

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

Abstract. We addressed the task to automatically recognize and normalize entities in a French medical corpus. To increase the coverage of our initial French terminology, English terms were translated into French by two different automatic translators. Indexing with a terminology that contained the intersection of the translated terms in combination with several post-processing steps to reduce the number of false-positive detections, gave the best performance results.

Keywords: Entity recognition, Concept identification, Term translation, French terminology

1 Introduction

The CLEF eHealth 2015 task 1b focuses on concept recognition in French medical text [1, 2]. The task consists of three subtasks: recognition of relevant entities in a French medical corpus, normalization of the recognized entities, and normalization of entity mentions that had been manually annotated. The entities covered a wide variety of semantic groups. The normalization had to be based on the Unified Medical Language System (UMLS), and involved assigning UMLS concept unique identifiers (CUIs) to the entities that were recognized or provided. Each subtask should be performed fully automatically.

We addressed all three subtasks. Central in our approach to entity recognition and normalization are French terminologies based on the UMLS and post-processing steps to reduce the number of false-positive detections. The UMLS already contains a number of French vocabularies, but their coverage is rather limited. We therefore explored the possibility to expand the coverage by automatic translation of English UMLS terms into French. For this purpose, we utilized two automatic translators.

2 Methods

2.1 Corpora

We utilized two corpora in our experiments: the Quaero medical corpus, a French annotated resource for medical entity recognition and normalization [3], which was the basis for the training and test sets provided in task 1b; and the Mantra corpus, a large multilingual biomedical corpus developed as part of the Mantra project [4], which we used to determine the terms for term translation and to create a term exclusion list. Each corpus is briefly described below.

Quaero Corpus. The Quaero corpus consists of three subcorpora (1): titles from French Medline abstracts, drug labels from the European Medicines Agency (EMEA), and patents from the European Patent Office. For the task 1b challenge, only Medline titles and EMEA documents were made available. The training set consisted of 833 Medline titles and 11 EMEA documents; the test set contained 832 Medline titles and 12 EMEA documents.

The annotations in the Quaero corpus are based on a subset of the Unified Medical Language System (UMLS) [5]. Briefly, the UMLS is a metathesaurus integrating more than 150 biomedical terminologies. Each concept in the UMLS is assigned a concept unique identifier (CUI), a set of corresponding terms, and one or more semantic types, which are mapped to one of 15 semantic groups (SGs) [6]. Typically, each concept belongs to one semantic group. An entity in the Quaero corpus was only annotated if the concept belonged to the UMLS and the corresponding SG was any of the following 10 SGs: 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).

Mantra Corpus. The Mantra corpus was developed as part of the Mantra project [4], aimed at providing multilingual resources in English, French, German, Spanish, and Dutch. The corpus consists of 1.6 million bilingual Medline titles (always in English and one of the other languages), 130k sentences of EMEA drug labels (available in all five languages), and 155k sentences of EPO patents (in English, French, and German in parallel). The texts in the Quaero corpus are a subset of the French texts in the Mantra corpus. The Mantra corpus is supplied with automatically generated silver-standard annotations, and recently multilingual gold-standard annotations have become available for a small subset of the Mantra corpus [7], but none of these resources were used in the current work.

2.2 Term Translation

The UMLS version 2014AB contains 178,860 unique French terms from 88,986 concepts, mainly stemming from MedDRA and MeSH, and only covering a few percent of the more than 5 million English terms and 2.6 million concepts in the UMLS. To expand the number of French terms, we used the web services application programming interface from Google Translate (GT) [8] and Microsoft Bing (MB) [9] to automatically translate English terms into French. Initially, we considered the translation of all English terms in the UMLS, but dismissed this approach as being too expensive and time-consuming. Instead, we reasoned that only the concepts that are found in a large English corpus that is representative of the task domain, may also be found in the Quaero corpus. We therefore indexed all English Medline titles and EMEA sentences from the Mantra corpus with our indexing system Peregrine [10], using the full English UMLS, and found 133,246 unique concepts. The 745,158 English terms corresponding with these concepts were translated into French using the automatic translators.

2.3 Terminologies

In our experiments on the Quaero corpus we used five French terminologies:

- Baseline: all French terms in UMLS version 2014AB. Only terms belonging to concepts in the ten SGs listed above were considered.
- GT: all terms from Google Translate and the baseline terminology.
- MB: all terms from Microsoft Bing and the baseline terminology.
- Union: all terms from Google Translate, Microsoft Bing, and the baseline terminology.
- Intersection: all terms that had the same translation by Google Translate and Microsoft Bing, supplemented with the baseline terminology.

The English terminology for indexing the Mantra corpus consisted of all English terms in UMLS version 2014 AB, filtered for the ten relevant SGs.

Both on the English terminology and the French baseline terminology we applied a set of term rewrite and suppression rules [11]. In a separate step (explained below), we supplemented the French terminologies with the concepts and terms in the training data.

2.4 Entity Recognition and Entity Normalization

The processing for the entity recognition and the entity normalization included an indexing and a post-processing step, which are described below.

Indexing. The corpora were indexed with Peregrine, a dictionary-based concept recognition system [10]. 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.

Post-processing. 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 Mantra corpus with the largest available French terminology (union), ordering the detected terms by their frequency in the corpus, 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 and had also been annotated in the training data, we computed precision scores: *true positives* / (*true positives* + *false negatives*). 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).

2.5 Normalization Based on Gold-Standard Entity Recognition

For entity normalization given the gold-standard terms and SGs, we developed the following processing pipeline. First, we computed precision scores for all term-SG-CUI combinations in the training set. If a given term-SG combination in the test set was also present in the training set, we selected the CUI of the term-SG-CUI combination with the highest precision score. If the second largest precision score was larger than 0.3, the CUI of the corresponding term-SG-CUI combination was also selected.

Second, if a term-SG combination in the test set had not been seen in the training set, we searched the terminology for terms that had a Levenshtein edit distance of maximum one. If one such term was found, the corresponding CUI was selected. If multiple terms were found, for each term the corresponding SG-CUI combination was sought in the training data. If present, precision scores were computed and the CUI of the SG-CUI combination with the largest precision was selected. If the SG-CUI combination did not exist in the training data, it was checked if the term was the preferred term for any of the CUIs. If this was the case for just one CUI, it was selected. Otherwise, a single CUI was selected, taking into account whether the CUI had been annotated in the training set, and the CUI number (lowest first).

3 Results

3.1 Performance on the Quaero Training Set

We used the Quaero training data to optimize the performance of the indexing and post-processing steps for entity recognition and normalization. Table 1 shows the results for the five French terminologies that we generated: Baseline (UMLS French), GT, MB, Union, and Intersection. The results have been generated with the task 1b evaluation script, using exact matching for both entity recognition and normalization.

Table 1. Performance of five French terminologies on the Quaero training set

		Entity recognition			Entity normalization		
Corpus	Terminology	Precision	Recall	F-score	Precision	Recall	F-score
EMEA	Baseline	0.724	0.399	0.515	0.588	0.359	0.446
	GT	0.368	0.763	0.496	0.220	0.670	0.332
	MB	0.345	0.791	0.481	0.208	0.687	0.316
	Union	0.298	0.807	0.435	0.172	0.702	0.274
	Intersection	0.454	0.756	0.567	0.273	0.669	0.388
Medline	Baseline	0.716	0.433	0.540	0.591	0.376	0.460
	GT	0.392	0.658	0.491	0.236	0.572	0.335
	MB	0.370	0.664	0.475	0.229	0.579	0.328
	Union	0.343	0.705	0.461	0.199	0.612	0.300
	Intersection	0.447	0.628	0.523	0.274	0.550	0.366

The terminologies based on automatic term translations (GT and MB) substantially increase recall as compared to the UMLS baseline terminology, but at the expense of a large decrease in precision. GT performs slightly better than MB in terms of F-score. The union of both terminologies results in a small further increase of the recall. The intersection improves precision considerably at the expense of some loss of recall. The performance of the terminologies with translated terms is better on the EMEA documents than on the Medline titles, primarily because the recall is higher. Interestingly, the reverse is true for the baseline terminology, which performs slightly better on the Medline titles. As expected, the performance for entity normalization is lower than for entity recognition, mainly because of a lower precision. This is largely caused by the ambiguity of many terms. At this stage, our indexing system did not try to disambiguate when multiple CUIs for the same term were found, and thus many of the CUIs were scored as false positives.

In our further experiments we decided to focus on the Union and Intersection terminologies. First, we tested the effect of expanding our terminologies with terms from concepts in the training data that were missed by our indexing system (false negatives). In order not to optimistically bias our performance results, we split the Quaero training data in an equally-sized training set and test set. Table 2 shows the performance results on the test set.

Table 2. Performance after expanding the terminologies with false negatives from half of the Quaero training set and testing on the other half

		Entity recognition			Entity normalization		
Corpus	Terminology	Precision	Recall	F-score	Precision	Recall	F-score
EMEA	Union	0.301	0.869	0.447	0.182	0.794	0.297
	Intersection	0.433	0.861	0.576	0.264	0.793	0.396
Medline	Union	0.401	0.708	0.512	0.246	0.638	0.355
	Intersection	0.513	0.668	0.580	0.326	0.607	0.424

Addition of the false negatives results in a clear improvement of the recall, with only a small decrease in precision.

Based on the expanded terminologies, we tested the effect of our post-processing steps, aimed at removing incorrectly indexed terms (false positives). An important parameter in this process is the precision threshold (see post-processing description above). Using half of the Quaero training data, we varied this threshold between 0.1 and 0.5 with steps of 0.1, and tested on the other half of the training data. The best F-score was obtained for a threshold of 0.3. Table 3 shows the results of the post-processing steps using this threshold.

Table 3. Performance of the expanded terminologies with post-processing steps on half of the Quaero training set

		Entity recognition			Entity normalization		
Corpus	Terminology	Precision	Recall	F-score	Precision	Recall	F-score
EMEA	Union	0.452	0.786	0.574	0.407	0.727	0.521
	Intersection	0.679	0.784	0.728	0.619	0.736	0.672
Medline	Union	0.579	0.605	0.592	0.477	0.508	0.492
	Intersection	0.747	0.581	0.654	0.634	0.500	0.559

The post-processing steps reduce recall but strongly increase precision, as well as the F-scores.

3.2 Performance on the Quaero Test Data

We submitted two runs for both the entity recognition and normalization tasks, one run using the Union terminology, the other using the Intersection terminology. Both terminologies were expanded with all false negatives of the Quaero training set. Table 4 shows our performance results on the final test set for exact match. (Note: we swapped the test run precision and recall values that the task organizers provided to us, since we could deduce from the FP and FN counts that they had been reversed.)

Our results on the test set were better than on the training set, mainly because of higher precision values. Overall, the system using the Intersection

Table 4. Entity recognition and normalization performance on the Quaero test set

		Entity recognition			Entity normalization		
Corpus	Terminology	Precision	Recall	F-score	Precision	Recall	F-score
EMEA	Union	0.710	0.776	0.741	0.653	0.705	0.678
	Intersection	0.751	0.761	0.756	0.707	0.714	0.711
Medline	Union	0.683	0.662	0.662	0.559	0.552	0.575
	Intersection	0.711	0.625	0.665	0.634	0.547	0.587

terminology performed best. These results are well above the average and median of the scores from all participant runs, as provided by the task organizers.

We also submitted two runs for the normalization using the gold-standard entity recognition results. The difference between the two runs was that the first run did not include the final disambiguation step (selection of CUIs if they had been annotated in the training set and based on CUI number). Table 5 gives the performance results.

Table 5. Normalization performance on the test set given the entity recognition, with and without the final disambiguation step

Corpus	Disambiguation	Precision	Recall	F-score
EMEA	No	1.000	0.767	0.868
	Yes	1.000	0.774	0.872
Medline	No	0.817	0.573	0.674
	Yes	0.805	0.575	0.671

As was to be expected, use of the gold-standard entity recognition improved the normalization results. In particular precision was boosted, with a remarkable precision of 1 for the EMEA corpus. The final disambiguation hardly affected the performance results.

4 Discussion

Our results show that expanding the coverage of the French UMLS baseline terminology with the use of an automated term translator is a viable way to improve the recall for entity recognition and normalization, but also reduces precision considerably. Taking the intersection of the term translations increases precision again, while only slightly reducing recall. The various post-processing steps further improve precision. The union of the term translations did hardly further improve the recall, indicating that the annotated corpus contained few terms that were uniquely provided by one of the translators. Although the precision of the Union terminology on the Quaero training set was substantially less than the precision of the Intersection, the difference on the test set was much smaller.

References

1. 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: CLEF 2015 – 6th Conference and Labs of the Evaluation Forum. Lecture Notes in Computer Science (LNCS). Springer, Heidelberg (2015)
2. 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)
3. 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)
4. Mantra project website, `http://www.mantra-project.eu`
5. Bodenreider, O.: The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic Acids Res. 32, D267–270 (2004)
6. Bodenreider, O., McCray, A.T.: Exploring Semantic Groups Through Visual Approaches. J. Biomed. Inform. 36, 414–432 (2003)
7. Kors, J.A., Clematide, S., Akhondi, S.A., van Mulligen, E.M., Rebholz-Schuhmann, D.: A Multilingual Gold-Standard Corpus for Biomedical Concept Recognition: the Mantra GSC. J. Am. Med. Inform. Assoc., epub ahead of print (2015)
8. Google Translate, `https://translate.google.com`
9. Microsoft Bing Translator, `http://www.bing.com/translator`
10. 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)
11. Hettne, K.M., van Mulligen, E.M., Schuemie, M.J., Schijvenaars, B.J.A., Kors, J.A.: Rewriting and Suppressing UMLS Terms for Improved Biomedical Term Identification. J. Biomed. Semantics 1, 5 (2010)