Multilingual Hybrid Automatic Term Extraction: the use case of ebpracticenet

Ayla Rigouts Terryn
Véronique Hoste
Joost Buysschaert
Els Lefever
LT3 Language and Translation Technology Team, Ghent University (Belgium)

Abstract

Accurate terminology is essential for professional communication, but also complex and challenging to translate. To improve multilingual communication, tools have been developed that automatically detect terms and their equivalents in other languages from parallel corpora. By means of a use case with data from ebpracticenet, we illustrate how hybrid multilingual automatic term extraction from parallel corpora works and how it can be used in a practical application such as search engine optimisation. The original aim was to use this list to improve the recall of a search engine by allowing multilingual searches (automatically obtaining search results containing both the original search term and the translations of the search term). Two additional possible applications were found when considering the data. The first addition was searching for related forms, using the automatically generated lemmas to group different forms of the same word. Next, it was found that multiple translations for the same source term reveal clusters of strongly semantically related words (e.g. the Dutch word "gif" is translated as "venom", "toxin" and "poison"), so these can be used to find relevant documents as well. The ebpracticenet use case clearly illustrates the practical use of automatic terminology extraction from parallel corpora and the benefits of real-world applications to provide inspiration for further research.

Keywords: automatic terminology extraction; ATR; terminology.

1. INTRODUCTION

Accurate and consistent terminology is essential for professional communication. This has led to the development of terminology management strategies, which often include tools to automatize different components of the terminology management workflow. This article is dedicated to the automatic extraction of multilingual terminology, using a hybrid approach, i.e., a combination of both linguistic and statistical features

to identify terminology. The practical use of this strategy will be illustrated by means of a use case for ebpracticenet, a Belgian digital database containing evidence-based medical guidelines and information for caregivers. The aim was to explore the possibilities of automatic term extraction (also known as automatic term recognition or ATR) for the optimisation of search engine recall. Multilingual ATR was performed on parallel corpora in English, French and Dutch. The acquired data inspired three different strategies for search engine optimisation. For each given search term, results can be found containing the search term itself, but also: (1) translations of the search term, i.e. documents about the same subject in a different language, (2) morphological variants of the search term, specifically terms with the same lemma, and (3) terms that are strongly semantically related to the search term. Additionally, auto-completion and suggestion of search terms can be improved with the monolingual lists of automatically extracted terms.

2. STATE-OF-THE-ART

ATR has been a productive field of research within computational linguistics. Early work often focussed on either linguistic (e.g. Bourigault, 1992), or statistical (e.g. Sparck Jones, 1972) clues to search for terms. Linguistically inspired methodologies rely on information such as part-of-speech patterns to identify terms, whereas statistical methods calculate word/term frequencies, often comparing frequencies in a specialised, domain-specific corpus, with frequencies in a large, general domain corpus. Kageura and Umino (1996) defined two of the fundamental concepts of automatic terminology extraction: termhood and unithood. Termhood refers to how characteristic or relevant a term is within the researched topic/domain. Unithood describes to which degree multi-word terms form a

syntagmatic linguistic unit. Since the linguistic and statistical approaches provide complementary information, later ATR methodologies (Daille, 1994) often combine the two approaches. These are called hybrid methodologies. Another evolution has been the introduction of a multilingual aspect by using parallel corpora to extract equivalents for terms in other languages as well. An example of a hybrid tool for bilingual ATR is TExSIS (Macken, Lefever, & Hoste, 2013), which was used for the experiments described in this article.

The evaluation of ATR has always been rather problematic due to the lack of an unambiguous definition of terms (Rigouts Terryn, Hoste, & Lefever, 2018). Terms are generally defined as lexical units which define relevant concepts within a specific domain. However, such definitions do not allow human annotators to identify terms without a certain measure of subjectivity. Consequently, inter-annotator agreement for term annotation is typically very low.

The two most important measures of ATR accuracy are precision and recall. Precision calculates how many of the automatically extracted candidate terms were evaluated as actual terms by human annotators. Recall measures how many of the terms found by human annotators in a text are also extracted automatically. While precision can be calculated based on the extracted list of terms, the calculation of recall necessitates a fully annotated corpus, large enough to be useful for ATR. Therefore, recall often is not calculated, especially for small-scale research.

3. TExSIS

The ATR tool used for this experiment is TExSIS (Macken, Lefever & Hoste, 2013), developed at Ghent University. TExSIS is a hybrid tool which can be used for both monolingual and bilingual ATR in English, French,

German and Dutch. Given a specialised, domain-specific corpus, TExSIS will, first, perform a shallow linguistic preprocessing, which includes automatic tokenisation, part-of-speech tagging and lemmatisation. The linguistic filter is rule-based and will extract all candidate-terms with predefined part-of-speech patterns (e.g. adjective+noun).

As the linguistic preprocessing alone overgenerates, candidate terms are put through several statistical filters. Termhood is measured by comparing relative frequencies of candidate terms in the specialised corpus with those in a large, general language corpus, using the term-weighting measure of Vintar (2010) and Log-Likelihood Ratio. C-value (Frantzi & Ananiadou, 1999) was chosen as a metric for unithood and for finding nested terms. The results are ranked based on Vintar's term weighting measure. For the experiment, the cut-off values at this stage were set very low to favour recall.

For multilingual ATR, TExSIS requires a sentence-aligned parallel input corpus. In that case, monolingual ATR will be performed on the two languages separately to generate two monolingual lists of term candidates. To generate translation suggestions for all candidate terms, automatic word alignment is performed, using GIZA++ (Och & Ney, 2003). Again, the decision was made to favour recall over precision for the translation suggestions.

4. USE CASE EBPRACTICENET

4.1. Data

Ebpracticenet[1] is a digital database of evidence-based medical guidelines and information for caregivers. The texts in this database are written in English, French and Dutch. While they contain a large number of aligned translations (i.e. parallel corpora) for English-French and English-Dutch,

only a limited portion of all texts in the database has been translated. Therefore, to improve the search engine recall, ebpracticenet wants to enable searching across the three languages, so that, for any given search term, the search engine can return both documents containing the search term and documents that contain a translation of the search term.

As input for TExSIS, ebpracticenet provided two sentence-aligned parallel corpora based on the translation of nearly one thousand medical guidelines. The source language was English for both corpora, the target languages French and Dutch respectively. The English-French corpus contains 1,101,217 tokens in English and 1,266,731 tokens in French. The English-Dutch corpus contains 1,147,311 English tokens and 1,137,773 tokens in Dutch.

4.2. Results from TExSIS

After running both corpora (English-French and English-Dutch) through TExSIS, English was used as a pivot-language to turn the two separate lists into three trilingual lists (one per language). Since not all candidate terms have a version in each language and some have multiple translations, three trilingual lists were made, each with a different language that acted as source language and the other two as target languages A and B. For each lemmatised candidate term in the source language, eight different types of information were given:

1. Possible translation in target language A (lemmatised candidate term). If no translation is available, "0" is added. Multiple translations can be listed, separated by a comma.
2. Possible translation in target language B.
3. List of full forms found in the corpus of the lemmatised source language candidate term.

4. Part-of-speech pattern of source language candidate term.
5. Frequency of source language candidate term (sum of frequencies of full forms).
6. Termhood score (averaged over all full forms).
7. Log-likelihood ratio (averaged over all full forms).
8. C-value (averaged over all full forms).

For instance, the English term "beta-blocker" is accompanied by the following information:

1. bêtabloquant, bêta-bloquant
2. bètablokker, beta-blokker, β-blokker
3. beta-blocker, beta-blockers
4. Noun
5. 202
6. 266
7. 2572
8. 0.14

Table 1 shows how many different lemmatised candidate terms were found for each language. By presenting all data in sortable tables, the cut-off values could be determined ad-hoc.

	EN	FR	NL
Lemmatised CTs with min. 1 translation	74,384	46,408	67,904
Lemmatised CTs with English translation	n.a.	46,408	67,904
Lemmatised CTs with French translation	45,512	n.a.	40,215
Lemmatised CTs with Dutch translation	64,113	37,012	n.a.

Table 1. Number of extracted lemmatised candidate terms (CTs)

Since English was used as a pivot language and French and Dutch corpora were not based on exactly the same English corpus, there are more

lemmatised candidate terms with one translation in English and all lemmatised candidate terms in French and Dutch have at least one English translation suggestion. The data revealed that only a small percentage of all lemmatised candidate terms appear with more than one full form in the corpus: 4-6%. However, since there are so many extracted terms, this still amounts to over ten thousand lemmatised candidate terms with multiple full forms in total.

To check the relevance of the data for the task, spot-checks were performed to calculate precision at different points in the ranked list (sorted on termhood score). These checks were performed on the English list. A candidate term was considered correct if (1) it was related to the medical domain and (2) could conceivably be used as a search term on the ebpracticenet website. To clarify: we did not evaluate termhood, but potential use as a search term in the ebpracticenet search engine. To compare accuracy in relation to rank (based on termhood measure), 50 terms were annotated at 7 different points: the first 50 terms, then 50 terms at 5%, 10%, 25%, 50% and 75% of the total ranking and the 50 bottom-ranked terms. In total, this resulted in annotations for 350 candidate terms. Inter-annotator agreement was calculated to ensure a nuanced interpretation of the results. The two annotators agreed 85% of the time, resulting in a Cohen's kappa score of 0.6.

	1%	**5%**	**10%**	**25%**	**50%**	**75%**	**99%**	**Total**
Validated	41	45	42	41	39	22	10	240
Discarded	9	5	5	5	5	24	27	80
Named Entity	0	0	3	4	6	4	13	30
Precision (incl. NEs)	82%	90%	90%	90%	78%	52%	46%	77%
Precision (excl. NEs)	82%	90%	84%	82%	90%	44%	20%	69%

Table 2. Precision at different termhood ranks (50 terms per percentile)

The results of the evaluation are presented in Table 3. First of all, we see a very high precision for the first half of the candidate terms, especially when including named entities. Even at the 75th and 99th percentile, up to half of the candidate terms could be relevant. The first explanation for the quality of these results is that we evaluated usefulness as search terms, not termhood. For instance, *insulin requirement of basal metabolism* could be used as a search term but, typically, *insulin requirement* and *basal metabolism* would be considered terms separately. The evaluation was also lenient by allowing relevant parts of potential search terms: e.g. *failure*, which can be combined in terms such as *organ failure* or *heart failure* but would not be considered a medical term on its own. Despite the limited scope of the evaluation, the results are convincing enough to indicate the practical use of ATR for search terms.

Precision was also calculated for the automatically generated translation suggestions. All previously validated terms were evaluated on the French and Dutch translation suggestions. Named entities and rejected terms were excluded from this analysis. Any translation suggestions that were equivalent in meaning to the source term or nearly so were validated. Translation suggestions of a different word class than the source term, but with the same general meaning were also validated (e.g. if the source term was *ill* (adjective), translations of *illness* (noun) were validated as well). Otherwise, the evaluation was very strict, discarding any hyponyms, hypernyms and other strongly related but not synonymous terms. In some cases, a French or Dutch text contained English terminology. These were also discarded, as well as any misspellings. The results of this analysis are presented in Table 4.

	FR	NL
# Validated English search terms with translation(s)	162	208
% of those with min. 1 correct translation	97%	98%
% of those with only correct translations	81%	82%
% of those with multiple correct translations	22%	24%
Average % of correct translations	89%	88%

Table 3. Precision of translation suggestions

Once again, the results look promising, with nearly all search terms having at least 1 good equivalent in the other languages. There were fewer equivalents in French, since that parallel corpus was slightly smaller, so some of the English terms simply did not occur in the English-French parallel corpus. A large proportion of all search terms have multiple translation suggestions, though not all of the suggested translations are correct. Highly ranked terms are often frequent terms, for which many potential translations are found. For instance, the term *disease* has 18 different translation suggestions in French and 26 in Dutch. While these lists contain correct translations (e.g. *maladie* in French and *ziekte* in Dutch), they also contain many incorrect suggestions. Rejected translations include the original English form *disease*, semantically related terms such as *problème/problem* (EN: *problem*) and *infection/infectie* (EN: *infection*), hyponyms such as the translations *dementia* and *lung infection*, and, in Dutch, there are also a few complex compound terms, which contain the correct translation, but only as part of the compound, e.g. *ziekteverloop* (EN: course of the illness). The previously cited example of beta-blockers reveals another type of related terms: different spellings (e.g. in Dutch *bètablokker*, *beta-blokker* and *β-blokker*). The more general the source term, the more diverse (and inaccurate) the translations. Rarer terms usually have only one,

often correct translation suggestion. Some of the most interesting cases are the specific, yet still frequent terms, such as *beta-blocker*. Translation suggestions for these terms are often lists of synonyms or alternative spellings, e.g. the translations for *cough medicine: antitussive* and *médicament contre le toux* (French) and *hoestmiddel, hoestmedicijn* and *hoestmedicatie* (Dutch).

4.3. Application: search and engine optimisation

As described, there are three possible applications of the data for the optimisation of search engine recall. For each given search term, the search engine should look, not only for the given search term, but also for:

- Translations of the search term
- Terms with the same lemma
- Strongly semantically related terms

The first application has already been implemented by Ebpracticenet, as shown in the screenshot in Figure 1, where the English search term *heart failure* also yields results in Dutch.

The other two applications have not been implemented yet, but the results above look encouraging for both. While only few terms have more than one full form per lemma, the automatically generated lemmas are often different from the full form and can also be linked. The third application will be most difficult since these data are less accurate. One possibility is to only use this strategy when the other ones yield no or very few results. This would automatically exclude the most frequent and general terms, for which there were many translation suggestions which were less accurate. Since the precision of TExSIS for the extraction of search terms is so good, an additional possible application presents itself, namely auto-completion of the

search term. The lowest ranked term candidates could be excluded to ensure a high precision.

Figure 1. Screenshot of multilingual search in ebpracticenet

While these results clearly indicate that multilingual ATR from parallel corpora is accurate enough to be usable without much additional human effort, two problems remain: (1) this method requires large, sentence-aligned parallel corpora, which are not always available and very expensive to create and (2) these corpora have to be carefully maintained to ensure the most recent terminology is included. To deal with this data acquisition bottleneck, research into multilingual ATR has recently focussed more on comparable corpora as a source of information.

In conclusion, (multilingual) automatic terminology extraction has evolved to become a useful tool in applications such as the improvement of search engine recall and one of the next steps will be to base extractions on comparable corpora.

NOTES

[1] https://www.ebpnet.be/

REFERENCES

Bourigault, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th conference on Computational linguistics-Volume 3* (pp. 977–981). Association for Computational Linguistics.

Daille, B. (1994). Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In J. Klavans & P. Resnik (Eds.), *The Balancing Act: Combining Symbolic and Statistical Aproaches to Language* (pp. 49–66). Massachusetts: MIT Press.

Frantzi, K. T., & Ananiadou, S. (1999). The C-value/NC-value domain-independent method for multi-word term extraction. *Journal of Natural Language Processing*, *6*(3), 145–179.

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition. *Terminology*, *3*(2), 259–289.

Macken, L., Lefever, E., & Hoste, V. (2013). TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment. *Terminology*, *19*(1), 1–30.

Och, F. J., & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, *29*(1), 19–51.

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2018). A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents. In *Proceedings of LREC 2018*. Miyazaki, Japan: ELRA.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–21.

Vintar, S. (2010). Bilingual Term Recognition Revisited. *Terminology*, *16*(2), 141–158.