Argentinian Journal of Applied Linguistics
Vol. 7, No. 1, May 2019, 93-108

FAAPI

Validating multilingual hybrid automatic term extraction for search engine optimisation: the use case of EBM-GUIDELINES

Ayla Rigouts Terryn* - Ghent University, Belgium
Véronique Hoste - Ghent University, Belgium
Joost Buysschaert - Ghent University, Belgium
Robert Vander Stichele - Ghent University, Belgium
Elise Van Campen - Editor, ebpracticenet, Belgium
Els Lefever - Ghent University, Belgium

(Received 20/12/18; final version received 01/04/19)

ABSTRACT

Tools that automatically extract terms and their equivalents in other languages from parallel corpora can contribute to multilingual professional communication in more than one way. By means of a use case with data from a medical web site with point of care evidence summaries (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 the result of the extraction to improve the recall of a search engine by allowing automated multilingual searches. Two additional possible applications were found while considering the data: searching via related forms and searching via strongly semantically related words. The second stage of this research was to find the most suitable format for the required manual validation of the raw extraction results and to compare the validation process when performed by a domain expert versus a terminologist.
Keywords: automatic terminology extraction; ATR; terminology.

RESUMEN

Las herramientas que extraen automáticamente términos y sus equivalentes en otros idiomas de corpus paralelos pueden contribuir a la comunicación profesional multilingüe de más de una manera. A través de un caso práctico con datos (extraídos de) ebpracticenet, ilustramos cómo funciona la extracción automática de términos multilingües híbridos a partir de corpus paralelos y cómo se puede utilizar en una aplicación práctica como la optimización de motores de búsqueda. El objetivo original era utilizar el resultado de la extracción para mejorar la recuperación de un motor de búsqueda permitiendo búsquedas multilingües automatizadas. Al considerar los datos, se encontraron dos posibles aplicaciones adicionales: la búsqueda a través de formularios relacionados y la búsqueda a través de palabras muy relacionadas semánticamente. La segunda etapa de esta investigación consistió en encontrar el formato más adecuado para la validación manual necesaria de los resultados de la extracción bruta y comparar el proceso de validación cuando lo realiza un experto en medicina frente a un terminólogo.
Palabras clave: extracción automática de terminología; ATR; terminología.

* Corresponding author e-mail: ayla.rigoutsterryn@ugent.be

ACCURATE AND CONSISTENT terminology is essential for professional communication. This has led to the development of terminology management strategies, which often include tools to automate different components of the terminology management workflow. This paper 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 EBM-GUIDELINES, a digital database with 1000 highly structured evidence-based guidelines (point of care evidence summaries), published by DUODECIM, the publishing company of the Finnish General Practitioners. All these guidelines have been translated to English and then to Dutch and French, to enable implementation in Belgium. 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, search engine results can be found containing the search term itself, and, in addition: (1) translations of the search term in different languages, (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 auto-suggestion of search terms can be improved with the monolingual lists of automatically extracted terms.

While the ATR method used reached a state-of-the-art performance, the results are not yet perfect and require manual validation before they can be implemented in a search engine. Before moving on to the validation, the data needed to be presented in a suitable format. With regard to terminological validation, there are two commonly used approaches for terminological validation: either the results are validated by a domain expert (without specific training in terminology), or they are validated by a terminologist (without domain expertise). In this case, a domain expert (a medical doctor) was consulted to validate the results of the multilingual term extraction. For this research project, a trained terminologist and translator also validated part of the data for comparison, with identical instructions. Since both validating terms and evaluating translations are known to be highly subjective tasks, it is interesting to consider the impact of the validator's background on this task.

The remainder of this paper is divided into four parts. First, the state-of-the art in the field of monolingual and bilingual term extraction is discussed in section 2. Section 3 describes the term extractor used for these experiments: TExSIS. Section 4 explains how these results might be used for search engine optimisation and includes a short evaluation of the results for that purpose. Section 5 is dedicated to the validation of the results, discussing both the methodology and a comparison of the results by the different annotators. Finally, the results are summarised and interpreted in the conclusion, along with suggestions for further research.

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 paper.

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 refer to relevant concepts within a specific domain. However, such definitions allow room for interpretation, so human annotators identify terms with 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. Both measures can be combined into f-score, which is the harmonic mean of precision and recall. Existing resources such as the IATE (Inter-Active Terminology for Europe) or MeSH (Medical Subject Heading) term banks can be used as a reference. For instance, Laroche and Langlais (2010) use 5000 nominal term pairs from MeSH. However, while using such established resources may decrease subjectivity and annotation effort, they do not reflect recall accurately, since there may always be valid and relevant terms in a corpus that are not present in a term bank. Moreover, since one of the applications of automatic term extraction is to extract new terms to keep these types of term banks up-to-date, this evaluation methodology may miss very relevant terms. Term Evaluator (Inkpen, Paribakht, Faez, & Amjadian, 2016) is a tool designed specifically to evaluate and compare different term extractors. The results of several tools are combined, and the tool provides an

interface in which to efficiently annotate the list of extracted term candidates. While this strategy does not allow the calculation of recall (since only the list of extracted term candidates is annotated, not the terms in the original texts), it does provide the option of calculating *relative recall*, taking the union of all term candidates, extracted by all term extractors, as an approximation of all possible term candidates in the text.

When it comes to the validation of term candidates, generally, a choice needs to be made about whether to have a domain specialist or a terminologist perform the annotation. Involving a domain specialist is not always an option, especially if multiple domains or languages are researched. While inter-annotator agreement scores are sometimes reported, the impact of the annotator on the term validation is rarely researched. Hätty and Schulte im Walde (2018) are a notable exception. They asked 20 laypeople to annotate terms in specialised texts in four different domains: do-it-yourself, cooking, hunting, and chess. The term identification was split into four tasks performed in WebAnno (Yimam, Gurevych, de Castilho & Biemann, 2013): highlighting domain-specific phrases, creating an index, defining unknown words for creating a translation lexicon and creating a glossary. There were seven annotators per task. The authors found that agreement was similar regardless of the task and that "laypeople generally share a common understanding of termhood and term association with domains", but that "laypeople's judgments deteriorate for specific and potentially unknown terms" (Hätty & Schulte im Walde, 2018, p. 325). In another study (Rigouts Terryn, Hoste, & Lefever, accepted), a terminologist annotating terms in different domains reported that, while annotating in a domain for which she was a domain specialist was faster, it can also be more difficult to recognise domain-specific terminology when that terminology has become part of one's general vocabulary.

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 term extraction from parallel corpora 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. Then, a rule-based linguistic filter extracts all candidate-terms with predefined part-of-speech patterns, both single words and multi-word units. Examples of patterns for English are: noun (e.g. *anaemia*), adjective+noun (e.g. *antiarrythmic* agent), noun+preposition+noun (e.g. *loss of consciousness*) etc. These patterns are, of course, language-dependent. One example of an important difference between languages here is the way compound terms are constructed. In Dutch, compound terms are usually single-word compounds, whereas in French and English, multi-word terms are more common. This directly influences the term extraction, due to the different strategies required for single-word or multi-word term extraction.

The linguistic preprocessing (i.e. identification of candidate terms based on POS-pattern) favours recall over precision, and hence generates too much terms. Therefore, candidate terms are put through a statistical filter. In this phase, several statistical scores are computed to calculate termhood and unithood. 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). Log-Likelihood Ratio (Rayson & Garside, 2000) is another such termhood measure, which is, in this case, only calculated for single-word terms. C-value (Frantzi & Ananiadou, 1999) was chosen to calculate unithood and for finding nested terms, by looking at the length and the relative frequency of the candidate term itself, versus that of all other candidate terms that enclose this candidate term. 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 identify equivalent terms in the parallel texts 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.

Besides the termhood (whether the term is relevant to the specialised domain) and unithood (whether separate words belong to a single unit. i.e. a multi-word term) measures, an additional statistic was added for the bilingual component of the ATR: FreqRatio. This metric compares the frequency of the source term candidate and the suggested target term candidate. The intuition behind this metric is, that equivalent terms will probably appear a similar number of times in a parallel corpus. FreqRatio expresses the relative difference in frequency between suggested equivalents and can be used as an additional filter. However, using a hard cut-off based on FreqRatio is not always recommended, since it is very sensitive to differences in frequency caused, e.g. by synonyms and variants.

Multilingual Automatic Term Extraction for EBM-GUIDELINES

Data

EBM-GUIDELINES is a digital database of evidence-based medical guidelines and information for caregivers. Originally in Finnish and English, the database has been translated in Dutch and French, for implementation in Belgium by the company IScientia, using augmented machine translation with a translation memory, and subsequent revision by a professional translator and a medical proof-reader (cf. Van de Velde et al., 2015). The database is accessible online to caregivers through the eHealth Platform and Internet (https://www.ebpnet.be/). An independent non-profit organisation, ebpracticenet, financed by the Belgian government, provides contextualisation of the information for the healthcare system. The texts in this database are written in English, French and Dutch. The EBM-

GUIDELINES are big parallel corpora, providing a large number of aligned translations for English-French and English-Dutch. In addition to the guidelines, the database also contains 5000 English-only summaries of systematic reviews. These reviews underpin recommendations within the guideline. They are based on the work of the Cochrane Collaboration, a worldwide network that specializes in the production and maintenance of systematic reviews of randomized clinical trials in the field of medicine (https://www.cochrane.org/ Last accessed on Dec 20, 2018). This is considered the nec plus ultra of evidence-based medicine.

Dutch-speaking users search the EBM-GUIDELINES using the Dutch interface and Dutch search terms. With the help of an alignment tool coupling search terms in English and Dutch, relevant results can be retrieved, not only from the EBM-GUIDELINES (in Dutch), but also from the English-only Evidence Summaries. Therefore, the term extraction was commissioned by ebpracticenet to improve the search engine recall, both for Dutch and French users. The alignment tool should enable searching across Dutch and English and French and English, 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, two sentence-aligned parallel corpora were provided with the translation of nearly one thousand medical guidelines, with an average of 4 pages and 100 aligned segments per document. The source language was English for both corpora, the target languages French and Dutch respectively. Not all English texts were translated in both target languages, so the two parallel corpora are very similar, but not identical in content. 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.

Results from TExSIS

After running TExSIS on both bilingual corpora (English-French and English-Dutch), English was used as a pivot language to create trilingual term lists for the preliminary evaluation. For instance, the English list was based on the English lemmatised candidate terms. Each row contained one lemmatised English candidate term (e.g. *aneurysm*), all full forms of that candidate term found in the corpus (e.g. *aneurysm* and *aneurysms*) and all possible translations of the (lemmatised) candidate term in French and Dutch, including all possible full forms of the suggested translations. Additionally, the information from the term extraction was added: the part-of-speech pattern, frequency and termhood and unithood scores. Separate lists were made based on the French and Dutch lemmatised candidate terms, since not all candidate terms have a version in each language, and some have multiple translations.

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. 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.

	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); n.a. = not applicable.

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. Moreover, these are often important and/or frequent terms, such as *patient*, *symptom* and other common medical occurrences such as *arrhythmia*, *haemorrhage* and *thrombosis*.

To check the relevance of the data for the improvement of the search engine, spot-checks were performed to calculate precision at different points in the ranked list (sorted on Vintar's 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 relevance as a search term in the ebpracticenet search engine. 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.

Evaluation of Results

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 termhood 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.

Termhood rank	1%	5%	10%	25%	50%	75%	99%	Total
Nr. of analysed terms	50	50	50	50	50	50	50	350
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. Even at the 75th and 99th percentile, up to half of the candidate terms could be relevant, especially when NEs are considered. The first explanation for the quality of these results is that we evaluated usefulness as search terms, not termhood. 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 selecting search terms.

Precision was also calculated for the automatically generated translation suggestions. All previously validated terms were evaluated with respect to the French and Dutch translation suggestions. Named entities and rejected terms were excluded from this analysis. All translation suggestions that were equivalent or nearly equivalent in meaning to the source term 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), OUT OF A TOTAL OF 350 CANDIDATE TERMS	162/350	208/350
% of validated terms with min. 1 correct translation	97%	98%
% of validated terms with only correct translations	81%	82%
% of validated terms 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 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. Translations that were judged as incorrect include the original English form *disease* (instead of a Dutch equivalent), semantically related, but non-equivalent terms such as *problème*/*problem* (EN: *problem*) and *infection/infectie* (EN: *infection*), hyponyms such as the translations *dementia* and *lung infection*. 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*). These were considered incorrect as well. The example of *beta-blockers* reveals another type of related terms: different spellings, e.g. in Dutch: *bètablokker*, *beta-blokker* and □*-blokker*. One more peculiarity we observed was, that, the more general the source term, the more diverse (and inaccurate) the translations. Rarer terms usually have only one, often correct translation suggestion. More general terms, such as *patient* or *disease,* appear very often (creating more room for mistakes) and are regularly translated less literally. For instance, a translator may choose to translate *patient* by *child* if, in a certain context, the two would clearly refer to the same person. In that case, TExSIS may, correctly, identify *child* as the translation, even though they are not equivalents in most cases. Finally, we also noticed how translation suggestions for these terms are often lists of synonyms or alternative spellings, e.g. the translations for *cough medicine*: *antitussive* and *médicament contre la toux* (French) and *hoestmiddel*, *hoestmedicijn* and *hoestmedicatie* (Dutch).

Application: Search Engine Optimisation

There are several ways in which these results, once validated, could contribute to an improved search engine. First and foremost, by allowing multilingual queries, e.g. where a search for *hartfalen* in Dutch would automatically search for *heart failure* in English as well. Second, variants of the same lemma can be searched, so that, e.g. *beta-blockers* would also return results for *beta-blocker*. The third and most difficult application would be to automatically look for strongly semantically related terms, which have the same translation. These data are less accurate, but may be worth considering for very common terms, such as *medicatie, geneesmiddel* and *medicijn* in Dutch. While we have not explored this option in any detail yet, our results do indicate that this may be an interesting next step. Finally, auto-completing terminology based on the known terminology could help users to formulate more relevant queries.

However, before any of these may be implemented, the results need to be manually validated to ensure reliable user-experience. The first step towards this goal is to present the data in a suitable format and to formulate strategies for efficient validation.

Validation by Domain Expert versus Terminologist

Format

The two main requests from ebpracticenet for the validation of the results were: (1) have one term candidate and equivalent suggestion per line and, (2) strategies to quickly eliminate/validate larger batches of term pairs. The former meant that three different bilingual lists needed to be created, each time choosing the *source* language. The three resulting lists are: English – French, English – Dutch, and French – Dutch. The results are based on the full form (not lemmatised) term candidates of the first (source) language. There is only one suggested equivalent in the target language per line, meaning that a single source term candidate may be repeated on several lines, once for each different suggested target language equivalent. The translations are linked based on the lemma, so if multiple full forms exist for a suggested equivalent, there will be multiple rows with one full form each. The latter requirement meant that the part-of-speech patterns, frequencies and all termhood and unithood measures were reported for both the source and target language term candidate, as well as the FreqRatio for the translation pair, as explained in section 3. The following list is an example of one row in the final English – Dutch table:

1. English full form: *beta blockers*
2. Dutch full form: *bètablokker*
3. English lemma: *beta blocker*
4. Dutch lemma: *bètablokker*
5. English POS: *singular noun*
6. Dutch POS: *singular noun*
7. Named Entity tags for the tokens of the English candidate term: *0 0*
8. Named Entity tags for the tokens of the Dutch candidate term: *0*
9. Length of English candidate term (in tokens): *2*
10. Length of Dutch candidate term (in tokens): *1*
11. Frequency of English candidate term: *36*
12. Frequency of Dutch candidate term: *179*
13. Vintar's termhood score for English candidate term: *12.5*
14. Vintar's termhood score for Dutch candidate term: *95.1*
15. C-Value for English candidate term: *35.*0
16. C-Value for Dutch candidate term: *0.25*
17. Log-likelihood ratio for English candidate term: *0*[1]
18. Log-likelihood ratio for Dutch candidate term: *1633*
19. FreqRatio: 397%

As can be seen in this example, the FreqRatio is quite high, even though the translation is correct. This is due to the fact that there are many different forms of this term in both languages and, in English, the term *beta blockers* appears more often with a hyphen: *beta-blockers*, while, in Dutch, the suggested form *bètablokkers* is the most common variant.

To automatically reduce the size of these lists before the manual validation process, a filter was created based on discussions with ebpracticenet about their preferences. Rather than simply filtering on, e.g. Vintar's termhood measure, the term length and frequency were also taken into account. Very long candidate terms with low termhood scores are rarely good terms and very infrequent terms with low termhood scores are rarely relevant. For instance, all terms with a termhood lower than one, were deleted. There were also filters that combined features, e.g. all terms where the product of the termhood score and the frequency was lower than 2.5 were discarded. These filters were determined experimentally and tuned so that, when applied to the English corpus, around 20k unique English lemmas remained. The table was accompanied by explanations about each column, including how they might be used to efficiently validate the results.

Annotation and Results

The actual validation was performed by a domain specialist (medical doctor), who is fluent in all three languages, but has no background in linguistics or terminology. For comparison, a trained translator and terminologist, fluent in all three languages, also performed a part of the validation. Both received the exact same instructions before the task and did not have access to each other's annotations. The instructions by ebpracticenet were not very specific. They wanted correct translation pairs of potentially relevant search terms to be used in their search engine and they wanted only translations in the same full form (e.g. for the English term *aetiology*, the Dutch term *etiologie* could be considered a good equivalent, but not the plural form *etiologieën*). While they hinted at wanting to make a glossary as well, the main purpose was to find relevant and correct translation pairs to improve the search engine. There were no specific instructions on how to deal with items like named entities, so the annotators developed their own strategies according to what they found logical. The annotations were only performed on the English-Dutch data. In total, a sample of 10,000 lines (with one English term candidate and one suggestion for a Dutch equivalent per line) was annotated by both annotators.

The resulting inter-annotator agreement is displayed in Table 5. In 88% of the cases, the annotators agreed, leaving 12% of the lines with different validations per annotator. Both annotators validated over half of the lines and the terminologist validated slightly more than the domain specialist. The resulting Cohen's kappa score for inter-annotator agreement is 0.75. Since evaluation of both terms and translations is notoriously difficult and subjective, this is a relatively high agreement.

	Domain specialist: valid	*Domain specialist: not valid*	*Total*
Terminologist: valid	4907	205	**5112**
Terminologist: not valid	1028	3860	**1233**
Total	**5935**	**4065**	**10000**

Table 5. Inter annotator agreement between terminologist and domain specialist

While annotating, the terminologist assigned the data into twelve different categories (see Table 6). Even though these categories are, of course, somewhat subjective, since they are based on the terminologist's assessment, they do allow for a more detailed analysis of the results. They also helped the terminologist to annotate more consistently and make the same decision for similar cases. Another difference in the annotation process between the terminologist and the domain specialist was, that the terminologist sorted the term candidates alphabetically (to easily group the same or similar terms and make a consistent decision) and the domain specialist sorted on termhood score (to prioritise the most relevant terms).

Annotation categories (by terminologist)	*Total # terms*	*Terminologist*	*Domain specialist:*	
			valid	*not valid*
Correct	5264	valid	4796	468
Incorrect	1606	not valid	9	1597
Same lemma, different full form	1685	not valid	16	1669
Incorrect but strongly related	175	not valid	4	171
Correct but number debatable	128	valid	4	124
Dutch = English	218	not valid	86	132
Not medical or relevant	369	not valid	87	282
Debatable	64	56 valid; 8 not	8	56
Named Entity: brand/medicine	21	valid	5	16
Named Entity: organisation	27	valid	17	10
Named Entity: person with initials	136	valid	51	85
Named Entity: person without initials	307	valid	29	278

Table 6. Categories of annotation as determined by terminologist

Based on the information in Table 6, the annotations were further analysed. Out of 5264 annotations which were considered correct by the terminologist, only 468 (±9%) were not validated by the domain specialist. In many of those cases, not annotating them was probably a simple result of human error. This is supported by the fact that, in at least 50 of these cases, a different full form of the same term pair was annotated as valid, e.g. the domain specialist annotated *bijwerking(en)* as a Dutch equivalent for *adverse event(s)* as valid in the plural form, but not in the singular. Similarly, the mistake may have been made on the side of the terminologist, causing more disagreement. Sometimes, the terminologists

could use her experience with terminology and translation to recognise less logical or more obscure translations, such as *knutten* as a translation for *biting midges*, *Alzheimer* as a correct translation for *Alzheimer's disease*, even without the explicit addition of a translation for *disease*, or recognising *spm* as a valid translation for *bpm* (abbreviation for *beats per minute)*. In other instances, the terminologist lacked the necessary domain expertise to easily recognise specialised terms, e.g. recognising that the Dutch term *sartanen* is a synonym for *angiotensine receptorblokkers* or knowing whether *arterial disease* can be translated as *vaatziekte* (literally *vascular disease*) or if these are different diseases.

In only nine cases did the domain specialist not agree on a term pair deemed incorrect by the terminologist, including small mistakes made by both annotators. More interesting categories are the next three: *same lemma different full form, incorrect but strongly related* (both semantically and morphologically, e.g. same concept but different part of speech) and *correct but number debatable* (e.g. when a singular term expresses the same meaning as a plural term in the other language). The task was to only consider term pairs correct when they are in the same form. Overall, the terminologist seems to have had the advantage in this case, annotating more consistently. Some of these differences are also due to how strict the instructions were interpreted. For instance, a term like *medication* (or *medicatie* in Dutch) is singular, so is it a correct translation for *drug* (*geneesmiddel)* but not for *drugs* (*geneesmiddelen*)? The domain specialist only validated the two singular forms, whereas the terminologist validated both. She reasoned that *medication* could be used to describe a collection of more than one *drug* and that, in the case of translations, *medication* could often be used for both the singular and plural forms. Similarly, for terms which can be used in plural but rarely are, e.g. *pain(s), discomfort(s), bleeding(s), tendency/-ies,* etc. The terminologist's strategy in these cases was to err on the side of leniency, while the domain specialist tended to only approve term pairs with the same grammatical number. This is also reflected in the category *debatable*, which contains term pairs that are perhaps not literal translations but could, in many cases, be used as equivalents. An example would be *oorzaak* (literally: *cause)* as an equivalent for *aetiological factor*. These two are not synonyms but, in the context of medical texts, can sometimes be used for the same concept. Another example is *coagulation*, which, technically, can refer to other concepts than *blood coagulation*, but in the context of these medical texts, it may be fair to assume they can be used as synonyms.

An exception to this pattern of leniency for the terminologist versus strictness of the domain specialist is in the case of untranslated (English) terms in the Dutch text. The terminologist only approved untranslated terms when they were Named Entities or so common in Dutch that they are used more or equally regularly than the actual Dutch term. For instance, the suggested Dutch translations for *ACE-inhibitor* were *ACE-remmer* or *ACE-inhibitor*. The latter variant may appear to be an untranslated English term at first sight, yet it is common enough to also appear with a Dutch plural form: *ACE-inhibitoren*, rather than the English plural *ACE-inhibitors*. This led the terminologist to accept both *ACE-remmer* and

ACE-inhibitor as valid Dutch translations. The domain specialist also rejects some very clearly untranslated terms but is less consistent when the difference is more subtle, e.g. an ending in -y instead of -ie, or a plural in -s instead of -en. The final categories to discuss are the Named Entities, for which we distinguish between organisations, brands and personal names. The terminologist decided to approve all of the above for the sake of consistency. The translations may not be very informative, since source and target terms should be identical, but the named entities can still be relevant search terms. The domain specialist generally rejected named entities (especially person names), but with many exceptions.

Overall, it appears that the annotators tackled this task with slightly different mindsets. The terminologist was less strict and more likely to approve non-literal translations than the domain specialist, who was stricter, except for non-translated (English) terms as Dutch equivalents. In conclusion, both the terminologist and the domain specialists had advantages and disadvantages. While the domain specialist was able to identify correct term pairs for specialized medical concepts more efficiently, the terminologist could use her experience to annotate more consistently and make informed decisions about term pairs which are less obviously equivalent. Some of these differences are, of course, due to the very general instructions for this task. Ideally, decisions such as whether to annotate proper names should be made beforehand by the client. Moreover, since only two annotators participated in this comparison, we should be careful about generalising these results. Still, the results suggest that the two annotation styles are complementary.

Conclusion

In this paper, we have shown how multilingual automatic term extraction from a parallel corpus has potential for a real-world application such as search engine optimisation. We showed how, despite the need for manual validation, ATE can efficiently produce a list of candidate terms that contains many relevant search terms and, for most of these, good equivalents are found automatically in the other languages. Four potential applications were suggested: (1) multilingual searches, (2) autocompletion of search terms, (3) searching for morphologically related forms using the automatic lemmatisation, and (4) searching for semantically related forms by clustering multiple translations (and back-translations) for the same candidate term.

A suitable format for validation was developed based on feedback from ebpracticenet, who are currently validating the dataset for the implementation of the first application. For this validation, they chose to collaborate with a domain specialist (a medical doctor), who is fluent in the three languages. This is a common strategy, since it is often assumed that domain expertise is necessary to efficiently manage terminology. However, we asked an experienced terminologist without domain expertise to also validate a sample of the dataset. This resulted in 10,000 shared annotations to compare. It was found that, while domain expertise can be an advantage in the case of very specialised terms, the experienced

terminologist was able to annotate more consistently. The domain specialist was generally stricter with the validation, but, since there were only two participants, it is unclear whether this was caused by the lack of detail in the instructions, which left room for the usual subjectivity of this task, or, whether these differences were due to the different backgrounds of the annotators. This would be an interesting path to investigate further, so that users may make a better informed and motivated decision about the person best suited for their validation task.

Even this small-scale study already suggests that, ideally, validation of the results of multilingual automatic term extraction for a real-world application such as search engine optimisation would happen in a multidisciplinary setting, i.e. involving both a terminologist and a domain specialist. Clear instructions should be determined beforehand, preferably combining the input of the client and both a terminologist and domain specialist. One strategy would be to start by having both annotators validate a small subsample and analysing the results and differences to formulate the most suitable strategy for the remainder of the task. Whichever strategy is preferred, the validation will likely benefit from the complementary skills of both a domain specialist and a terminologist.

Notes

1. Since log-likelihood ratio is only calculated for single-word terms, the English multi-word term receives a score of zero.

References

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

Carroll, J. M., & Roeloffs, R. (1969). Computer selection of keywords using word-frequency analysis. *Journal of the Association for Information Science and Technology*, *20*(3), 227–233.

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.

Hätty, A., & Schulte im Walde, S. (2018). A Laypeople Study on Terminology Identification across Domains and Task Definitions. *Proceedings of NAACL-HLT 2018*, 321–326. New Orleans, USA: ACL.

Inkpen, D., Paribakht, T. S., Faez, F., & Amjadian, E. (2016). Term Evaluator: A Tool for Terminology Annotation and Evaluation. *International Journal of Computational Linguistics and Applications*, *7*(2), 145–165.

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

Laroche, A., & Langlais, P. (2010). Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 617–625. Beijing, China.

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.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora, 38th Annual Meeting of the Association for Computational Linguistics*, 1–6. Hong Kong, China: Association for Computational Linguistics.

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

Rigouts Terryn, A., Hoste, V., & Lefever, E. (2019). In No Uncertain Terms: A Dataset for Monolingual and Multilingual Automatic Term Extraction from Comparable Corpora. *Language Resources and Evaluation*, 1–34. https://doi.org/10.1007/s10579-019-09453-9

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

Van de Velde, S., Macken, L., Vanneste, K., Goossens, M., Vanschoenbeek, J., Aertgeerts, B., Vanopstal, K., et al. (2015). Technology for large-scale translation of clinical practice guidelines : a pilot study of the performance of a hybrid human and computer-assisted approach. *JMIR Medical Informatics*, 3(4).

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

Yimam, S. M., Gurevych, I., de Castilho, R. E., & Biemann, C. (2013). WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 1–6. Sofia, Bulgaria: ACL.