Multi-word term extraction from comparable corpora by combining contextual and constituent clues

Nikola Ljubešić[1], **Špela Vintar**[2], **Darja Fišer**[2]

[1]Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb
[2]Department of Translation, Faculty of Arts, University of Ljubljana

Abstract

In this paper we present an approach to automatically extract and align multi-word terms from an English-Slovene comparable health corpus. First, the terms are extracted from the corpus for each language separately using a list of user-adjustable morphosyntactic patterns and a term weighting measure. Then, the extracted terms are aligned in a bag-of-equivalents fashion with a seed bilingual lexicon. In the extension of the approach we also show that the small general seed lexicon can be enriched with domain-specific vocabulary by harvesting it directly from the comparable corpus, which significantly improves the results of multi-word term mapping. While most previous efforts in bilingual lexicon extraction from comparable corpora have focused on mapping of single words, the proposed technique successfully augments them in that it is able to deal with multi-word terms as well. Since the proposed approach requires minimal knowledge resources, it is easily adaptable for a new language pair or domain, which is one of its biggest advantages.

1. Introduction

Resource-poor language pairs and domains can benefit greatly from the increasingly popular field of bilingual lexicon extraction from comparable corpora. The approaches bootstrap lexica of general as well as domain-specific vocabulary from large, usually web-based collections of texts in two languages that are not translations of each other but rather share common properties, such as subject field, time of publication, target audience etc.

Term extraction from comparable corpora is usually understood as a task that combines monolingual term recognition in each of the languages and cross-lingual term alignment using various techniques. Fung and McKeown (1997) and Rapp (1995) are considered the beginners of the alignment approach based on the hypothesis that two terms are likely to be translations of each other if they occur in similar contexts. Several authors experiment with different measures of context similarity (Chiao and Zweigenbaum, 2002; Morin et al., 2007) and report up to 80% accuracy in finding the correct translation among the 20 best candidates. Some approaches extend the bilingual mapping through cognate detection (Saralegi et al., 2008), while Lee et al. (2010) propose an EM-based hybrid model for term alignment.

It should be noted that these early approaches deal almost exclusively with single-word terms, and also that nearly all authors conclude that the size and comparability of the corpora play a key role in achieving good performance. In our previous work we too have shown a strong positive correlation of the degree of corpus comparability and size (Ljubešić et al., 2011). In addition, we have established that good coverage of the seed lexicon that is used to translate the features in the context vectors plays a much bigger role than its size, and that the seed lexicon can be built completely automatically provided that there is a lexical overlap between two closely-related languages (Fišer and Ljubešić, 2011). However, in all our previous experiments in lexicon extraction, we, just like most related work, have not tackled multi-word expressions, which are very important in natural language processing and for which there are even fewer already existing resources, especially because a number of domains evolve at a great speed, making the static resources obsolete very quickly.

The bag-of-equivalents term alignment approach is an effective method of finding multi-word-to-multi-word term equivalents. It is similar to the compositional approach used by Morin and Daille (2010) or to the abduction method described by Carl et al. (2004), however both of the above use predefined lexico-syntactic patterns to predict term variations. Our approach is more robust, however it requires a domain-specific translation lexicon, ideally with several translation possibilities, and this may not be readily available (Vintar, 2010). The main goal of this paper is to show that by enriching the lexicon with automatically extracted domain-specific single-word terms the overall performance of multi-word term extraction from a comparable corpus can be significantly improved.

This paper is structured as follows: in the next section we present all the resources and tools that were used in the experiment. The experimental setup is described in detail in Section 3. The results are evaluated and discussed in Section 4, and the paper is concluded with some final remarks and ideas for future work.

2. Resources and tools used

2.1. Comparable corpus

The main source of lexical knowledge in this experiment was the English-Slovene comparable corpus of on-line articles on health and lifestyle, which had already been used successfully in our previous research (Fišer et al., 2011). Health-related documents were extracted from the ukWaC (Baroni et al., 2009) and slWaC (Ljubešić and Erjavec, 2011) web corpora by the criterion that the cosine similarity to a domain model had to be higher than 0.25. The domain model was built on documents from the two main health-related Internet domains. It is based on content words as features and TF-IDF feature weights where the IDF weights were calculated on a news-domain corpus.

The subset of the constructed domain corpus we used in this experiment contains 1.5 million tokens for each language.

2.2. Seed lexicon

The seed lexicon used as an anchor between the two languages was constructed from the freely available Slovene-English and English-Slovene Wiktionaries that cover mostly general vocabulary. The entries from both Wiktionaries were merged and if the same pair of words was found in both resources, they were given a higher probability. The seed lexicon constructed in this way contains 6.094 entries.

2.3. LUIZ

LUIZ is a hybrid bilingual term extractor that uses parallel or comparable corpora as input and outputs mono- and bilingual lists of term candidates (Vintar, 2010).
Term recognition is performed on the basis of user-adjustable morphosyntactic patterns provided for each language. Then the extracted candidate phrases are assigned a term-hood value by comparing the frequency of each word to a reference corpus. Term alignment is performed using the bag-of-equivalents approach (Vintar, 2010), which presupposes a probabilistic bilingual lexicon as input. A list of possible translation candidates for a source multiword term is proposed by comparing each target term candidate to a bag of potential translation equivalents provided by the lexicon and computing an equivalence score.

2.4. ccExtractor

ccExtractor is a context-based bilingual lexicon extraction tool that was built during our previous experiments (Ljubešić et al., 2011; Fišer et al., 2011; Ljubešić and Fišer, 2011). It consists of a series of scripts that enable:

- building context vectors for a list of headwords from each corpus,
- translating features of context vectors from source language to target language via an existing seed lexicon and
- calculating the best translation candidates between headwords in the source language and the target language.

In this research the tool is used to enhance the general small seed lexicon used for multi-word term alignment with LUIZ.

3. Experimental setup

The main task in the experiment was to find translation candidates for multi-word terms from the health comparable corpus. In order to achieve this, the experiment was divided into three parts.
In the first part of the experiment we used LUIZ to extract multiword term candidates from both corpora. The result is a list of 25,865 English and 27,102 Slovene multiword term candidates.
In the second part of the experiment we aligned the extracted multiword term candidates between English and Slovene with LUIZ via our seed lexicon.
In the third part of the experiment we tried to improve the results by enhancing the seed lexicon used by LUIZ with 412 translation equivalents of the domain-specific vocabulary in the corpus that is not covered in the seed lexicon, which we obtained with ccExtractor. Term extraction and alignment were then repeated with the same settings, the only difference being the extended seed lexicon.
With this step we combined contextual information obtained from ccExtractor with the constituent information provided by LUIZ.

3.1. Term extraction

Term recognition in each part of the corpus was performed with the help of a predefined set of morphosyntactic patterns for each language. These patterns describe part-of-speech sequences of mainly noun phrases up to 5 words in length. Once candidate phrases were extracted from the corpora, a term weighting measure was used to assign a termhood value to each phrase. This measure computes single-word termhood by comparing the frequency of each word ($f_{n,D}$) to a reference, non-specialized corpus ($f_{n,R}$), and then combines the termhood scores of all constituent words with the frequency (f_a) and length (n) of the entire candidate phrase.

$$W(a) = \frac{f_a^2}{n} * \sum_{1}^{n} (\log \frac{f_{n,D}}{N_D} - \log \frac{F_{n,R}}{N_R}) \qquad (1)$$

3.2. Term alignment

The extracted multi-word terms were then aligned in the bag-of-equivalents fashion (see section 2.3) using the seed bilingual lexicon. For a given source multi-word term each target term candidate is compared to a bag of potential translation equivalents provided by the lexicon and an equivalence score is computed, thus generating a ranked list of possible translation candidates. If, for example, the bilingual lexicon contains the English-Slovene entries

```
blood  kri     1.0
flow   pretok  0.66  tok  0.33
```

the bag-of-equivalents for the English term candidate *blood flow* will contain all three equivalents, *kri*, *pretok* and *tok*. We now compare the Slovene term candidates to the bag and compute the equivalence score as the sum of the translation probabilities found in the target term, normalized by term length. Thus, for the above English term we extract

```
pretok krvi      0.83
tok krvi         0.66
šibak tok krvi   0.43
```

This approach is able to identify several good translation equivalents for a source term, which is especially valuable in domains with less standardized terminology and a lot of term variation. Furthermore, this approach is also able to find translation equivalents for the terms for which seed lexicon entries are missing or faulty.
In our current setting we are able to identify multi-word-to-multi-word equivalents of different lengths, but we do not identify single-word-to-multi-word term pairs.

3.3. Extension of the seed lexicon

In the third part of the experiment the idea was to extend the alignment of the extracted multi-word terms with the extension of the seed lexicon by adding the most relevant vocabulary from the corpus. Using the ccExtractor, we extracted three most probable Slovene translations for all English lemmas that were not already included in the initial seed lexicon.

The headwords in both parts of the corpus had to satisfy the minimum frequency constraint of 50 occurrences which is the most reasonable frequency threshold as proven in our previous experiments (Ljubešić et al., 2011). When building context vectors, a window of three lemmas on both sides of the headword was used and the collected features were weighted by the TF-IDF score. Context similarity was calculated with the Dice similarity metric. The probabilities of the translation candidates were calculated as their context similarity weights scaled to a probability distribution.

There were 412 English lemmas in the corpus that had not been present in the seed lexicon already and that satisfied the occurrence frequency criterion. Therefore, our extended seed lexicon contains 6.506 entries. This lexicon was used in the second run of the experiment in which all the other settings were the same as in the first run.

4. Evaluation of the results

In this section we report the results of manual evaluation of term extraction in both languages as well as the quality of term alignment. We focus here on measuring the accuracy of term extraction and alignment and while recall would be interesting to study more closely as well, we were not able to do it in this experiment because in order to measure it, we would need either a comprehensive terminological dictionary of this area for measuring absolute recall or a manually annotated corpus with multi-word terms in both languages for measuring recall relative to the terms used in the corpus.

4.1. Evaluation of term extraction

In total, 25,865 term candidates were extracted from the English part of the corpus and 27,102 from the Slovene part. The extracted term candidates were assigned a termhood score and in order to evaluate the quality of the extracted terms, we manually evaluated 100 highest-ranked term candidates for each language.

In the evaluation scheme, each candidate was categorized into one of three possible categories:

- the candidate was a correctly extracted multi-word term from the health domain;
- the candidate was a correctly extracted multi-word term but did not belong to the health domain;
- the candidate was not correctly extracted (a part of a multi-word term) or the multi-word expression was not a term.

The results of manual evaluation are shown in Table 1. Among the English candidates, 76 were correctly extracted health terms (e.g. *blood test*), 5 were terms but belonged to some other domain (e.g. *primary school*) and 19 of the candidates were either incorrectly extracted multi-word terms or multi-word expressions that belong to the general vocabulary (e.g. *next year*). The results for Slovene are slightly better: 86 of the candidates were correct, 3 were terms from a different domain and 11 were incorrectly extracted multi-word terms or other multi-word combinations. The reason for better results in Slovene is probably a cleaner, less noisy corpus, both in terms of domain-specific documents and in terms of corpus annotation because slWaC was built much more conservatively than ukWaC.

Term quality	English	Slovene
good term	76%	86%
term from a different domain	5%	3%
not a term	19%	11%

Table 1: Evaluation of term extraction on 100 highest ranked term candidates

An interesting characteristic in the highest-ranking term candidates is their length. In both languages, two-word terms are by far the most frequent, with only 4 English and 6 Slovene candidates that are longer than two words. On the one hand, this is to be expected because the longer the term, the less frequent it is in the corpus. But it also must be noted that the corpus does not contain expert medical texts but mostly magazine articles with health issues and lifestyle advice for the general public that contain fewer complex medical terms.

4.2. Evaluation of term alignment

The quality of term alignment was evaluated for each run of the experiment, with the original and the extended seed lexicon, in order to evaluate the impact of seed lexicon extension.

The extension of the seed lexicon was evaluated in our previous work (Fišer et al., 2011). It has a correct translation in the first position in 45% of cases while in additional 11% of cases there is a correct translation among the first ten candidates. We did not measure specifically the percentage of correct translations on the first three positions used in this research.

In this part of evaluation we checked the proposed term pairs and measured the accuracy of term alignment by manually inspecting the list of 477 multi-word term pairs that received an equivalence score higher than 0.5 in either run of the experiment. In the list 380 of these pairs were identical in both runs of the experiment while translation suggestions for 97 of the source terms were different with the two different seed lexicons. First we evaluate the termhood of the source language candidates and then, in case the candidates are considered a term, we evaluate the accuracy of its translation.

The evaluation schema used when evaluating termhood is:

- good term;
- term from a different domain;

- not a term,

while the evaluation schema used for evaluating the translation quality is:

- correct translation;
- close translation;
- incorrect translation.

Score	Percentage
good term	43.6%
term from a different domain	12.6%
not a term	43.8%

Table 2: Evaluation of term extraction on the 477 source language term candidates with equivalence score higher than 0.5

As Table 2 shows, source language term candidates that have good probable translation equivalents (equivalence score higher than 0.5) are partial or full terms in 56% of the cases. This is much lower than when evaluating the top ranked term candidates. In our opinion, there are two reasons for that:

- these are the terms with a high equivalence score, not a high termhood score;
- term candidates with a high equivalence score consist of constituents found in the general seed lexicon from which terms are rarely built.

The quality of term alignment is shown in Figure 1. We stress once again that term alignment evaluation was performed only on those pairs that were good terms in the source language. When using the original seed lexicon, translations for 41.5% of the terms are correct or close to correct, while, when using the extended seed lexicon, 52.2% of translations are correct or close to correct. It is interesting to note that there is an increase of almost 8% of the correctly aligned terms while the number of close to correct terms goes up by 3%. At the same time, the number of incorrectly aligned terms goes down by almost 11%. This can be considered a very big improvement and clearly shows that it is very beneficial to add the most relevant vocabulary for the particular domain or corpus to the seed lexicon, even if the equivalents are extracted automatically and are therefore somewhat noisy.
Another interesting observation is the fact that the pairs that were shared among the two seed lexicons are of a relatively high quality already and that the extension of the seed lexicon helped in exactly those cases that the original lexicon was not able to handle well at all, either because it was too small in size or too general for this particular domain. This shows that the already existing resources can easily and successfully be complemented with a simple and fully automatic technique such as ours, giving a big boost to the quality of term alignment.

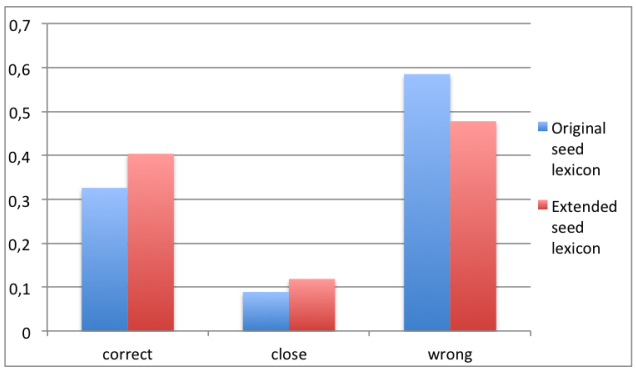

Figure 1: Evaluation of term alignment on terms with the equivalence score higher than 0.5 using the original and extended seed lexicon

5. Conclusions and future work

In this paper we presented an approach to extract translations of multi-word terms from domain-specific comparable corpora, a problem which has so far been largely neglected by most of the related work. We used LUIZ, a hybrid tool for bilingual multi-word term extraction and alignment. In addition, we used ccExtractor, a statistical tool for finding translation equivalents for single-word terms in comparable corpora in order to extend the seed lexicon with the most relevant terms in the corpus, which improved the results of multi-word term alignment by almost 11%. Additionally, this is the first extrinsic evaluation of context-based single-word lexicon extraction from comparable corpora.
While these results do not outperform the benchmark results achieved by LUIZ when aligning multi-word terms in parallel corpora, this is understandable because looking for MWT equivalents in comparable corpora is a much more difficult task. In addition, although the number of resulting MWTs obtained in this experimental setting is not very large, their precision is much higher than in the regular SWT extraction and alignment approach. With this in mind, the results we obtained with the extended seed lexicon are very encouraging and can already be very useful as a time-saving aid to terminologists who no longer have to look for the terms and their equivalents themselves but merely validate/correct the proposed ones.
Further improvements are possible by increasing the corpus size, which would, to start with, yield more single-word term candidates. This would improve the coverage of MWTs but could possibly have an adverse effect as well if a larger amount of noisy data in the lexicon would decrease the precision of the alignment. Finally, the term extraction procedure would benefit from more data as well.
In the future we plan to use the approach on a more scientifically-oriented medical domain corpus where complex terms play an even bigger role and there is less general language. Currently, we are also working on the adaptation of LUIZ to handle new languages, such as Croatian, which will enable the creation of multilingual terminological resources from web-based domain-specific comparable corpora.

6. Acknowledgement

Research reported in this paper has been supported by the ACCURAT project within the EU 7th Framework Programme (FP7/2007-2013), grant agreement no. 248347, and by the Slovene national postdoctoral grant no. Z6-3668.

7. References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, pages 209–226.

Michael Carl, Ecaterina Rascu, and Johann Haller. 2004. Using weighted abduction to align term variant translations in bilingual texts. In *Proceedings of LREC 2004*.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th international conference on Computational linguistics - Volume 2*, COLING '02, pages 1–5, Stroudsburg, PA, USA. Association for Computational Linguistics.

Darja Fišer and Nikola Ljubešić. 2011. Bilingual lexicon extraction from comparable corpora for closely related languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 125–131, Hissar, Bulgaria. RANLP 2011 Organising Committee.

Darja Fišer, Nikola Ljubešić, Špela Vintar, and Senja Pollak. 2011. Building and using comparable corpora for domain-specific bilingual lexicon extraction. In *4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 19–26, Portland. Association for Computational Linguistics.

P. Fung and K. McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, pages 192–202.

Lianhau Lee, Aiti Aw, Min Zhang, and Haizhou Li. 2010. Em-based hybrid model for bilingual terminology extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 639–646, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrwac and slwac: Compiling web corpora for croatian and slovene. In Ivan Habernal and Václav Matousek, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, volume 6836 of *Lecture Notes in Computer Science*, pages 395–402. Springer.

Nikola Ljubešić and Darja Fišer. 2011. Bootstrapping bilingual lexicons from comparable corpora for closely related languages. In *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, volume 6836 of *Lecture Notes in Computer Science*, pages 91–98. Springer.

Nikola Ljubešić, Darja Fišer, Špela Vintar, and Senja Pollak. 2011. Bilingual lexicon extraction from comparable corpora: A comparative study. In *First International Workshop on Lexical Resources, An ESSLLI 2011 Workshop, Ljubljana, Slovenia - August 1-5, 2011*.

Emmanuel Morin and Béatrice Daille. 2010. Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, 44(1):79–95.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 664–671, Prague, Czech Republic, June. Association for Computational Linguistics.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Meeting of the Association for Computational Linguistics*, pages 320–322, Cambridge, MA, USA.

X. Saralegi, I. San Vicente, and A. Gurrutxaga. 2008. Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. In *6th International Conference on Language Resources and Evaluations (LREC'08) - Building and using Comparable Corpora workshop*, pages 27–32, Marrakech, Morocco.

Špela Vintar. 2010. Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16(2):141—158.