Bilingual Terminology Extraction in Sketch Engine

Vít Baisa[1,2], Barbora Ulipová, Michal Cukr[1,2]

[1] Natural Language Processing Centre,
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
baisa@fi.muni.cz
[2] Lexical Computing
Brighton, United Kingdom and Brno, Czech Republic
{vit.baisa,michal.cukr}@sketchengine.co.uk

Abstract. We present a method of bilingual terminology extraction from parallel corpora and a few heuristics and experiments with improving the performance of the basic variant of the method. An evaluation is given using a small gold standard manually prepared for English-Czech language pair from DGT translation memory [1]. The bilingual terminology extraction (ABTE[3]) is available for several languages in Sketch Engine—the corpus management tool [2].

Keywords: terminology extraction, bilingual terminology extraction, Sketch Engine, logDice, parallel corpus

1 Introduction

Parallel corpora are valuable resources for machine and computer-assisted translation. Here we explore a possibility of extracting bilingual terminology from parallel corpora combining a monolingual terminology extraction [3] and a co-occurrence statistics [4]. We describe the method and how it is incorporated in the corpus manager tool Sketch Engine. We experimented with parameter tuning and evaluated a few settings using a small gold standard for English-Czech language pair.

The following section is a brief survey of topics, methods and tools in ABTE. In Section 3 we describe the basic algorithm for the extraction and in Section 4 how it is integrated in Sketch Engine. In Sections 5 and 6 we evaluate the algorithm and its variants.

2 Related work

The monolingual terminology extraction is a well-studied field, and the topic of ABTE has been explored since 90s [5] but recent summarizing publication [6]

[3] ATE stands for "automatic terminology extraction", so we adopt the abbreviation here and add "B" for "bilingual".

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2015, pp. 61–67, 2015.

do not mention this area of research: it is not yet a well-established area of the terminology field. It is probably partially caused by the fact that the quality of ABTE methods is rather poor.

ABTE functions are available in several commercial tools, e.g. MultiTerm[4], Araya Bilingual Extraction Tool[5], but it is hard to find any particular numbers referring to the quality of these tools.

Several ABTE-related papers can be split into two groups: 1) those using a combination of a monolingual extraction together with a co-occurrence statistics (e.g. [5], our algorithm belongs here too) and 2) those implementing an alternative approach.

An example of an alternative approach is described in [7] where authors used a bootstrapping technique: they started with discovering confident terminology pairs and then extracted rules for correspondence between terms in different languages (e.g. French-Dutch rule N+ADJ $\leftrightarrow$ ADJ+N).

Another approach is to use the phrase-based translation model approach as authors of [8]. According to [7], multi word:single word terms ratio is 7:3[6] so it is not viable to resort to single word alignment methods. It is though possible to start with a word alignment (using e.g. the Expectation-maximization algorithm) and then to extract pairs of phrases which are consistent with the word alignment.

3 The algorithm

Technically, parallel corpora in Sketch Engine are stored as monolingual corpora. The parallel alignment (usually on the sentence level) is defined by a mapping of IDs of special `<align>` structures or IDs of sentences (alternatively paragraphs and documents). The former is used when a 1:1 alignment is available (e.g. when working with TMX files, which is the case for DGT translation memory [1]). The latter is used in all other cases (e.g. in OPUS2 [9], EuroParl [10]).

The process of extraction of parallel terms consists of several steps. The first step is the monolingual terminology extraction in both languages. The procedure is described in [3]. The important thing to mention here is that all term candidates are found by means of matching grammar rules defined with CQL[7] (corpus query language) which are usually matching noun phrases in various forms.[8] For the ABTE method described here, it is not necessary to sort the (monolingual) term candidates by termhood by comparing term candidate frequency in a focus corpus with a reference corpus [3].

In a next step, the algorithm computes co-occurrence statistics for all aligned structures and for all candidate pairs occurring within the aligned structures.

[4] `www.sdl.com/cxc/language/terminology-management/multiterm`

[5] `www.heartsome.de/en/termextraction.php`

[6] It was measured in the domain of automotive industry.

[7] `www.sketchengine.co.uk/corpus-querying`

[8] According to [7], the majority of terms is in the form of noun phrases.

The resulting list of all term pairs can be sorted by various scores. By default Sketch Engine uses logDice[9] [?].

4 Bilingual terminology extraction in Sketch Engine

Bilingual terminology extraction is one part of a complex Sketch Engine's pipeline for building parallel corpora. For ABTE to work, it is necessary to have the monolingual terminology extracted for all languages in the relevant language pairs. Supported languages are: Chinese, Czech, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish.

L1 term	L2 term	Logdice	Co-freq	L1 freq	L2 freq
prevalence	prévalence	-0.0257005103	306	316	307
soap	savon	-0.0580571016	207	220	211
survival	survie	-0.0683060134	165	170	176
education	éducation	-0.0705785710	1815	1968	1844
adolescence	adolescence	-0.0711610289	89	91	96
condom	préservatif	-0.0840642648	125	139	126
primary prevention	prévention primaire	-0.0840642648	25	27	26
chronological age	âge chronologique	-0.0848888976	33	36	34
basic information	informations de base	-0.0874628413	16	17	17
acid	acide	-0.0874628413	16	17	17
rotavirus	rotavirus	-0.0931094044	15	16	16
universal access	accès universel	-0.0981803939	142	151	153
international guidance	directives internationales	-0.0995356736	14	15	15
stigma	stigmatisation	-0.1040724541	127	133	140
fish	poisson	-0.1043366598	20	21	22
pregnancy	grossesse	-0.1059334447	210	230	222
alcohol	alcool	-0.1110313124	25	28	26
vol	vol	-0.1168136650	83	87	93
syphilis	syphilis	-0.1233824155	28	32	29
public health	santé publique	-0.1235746851	123	133	135

Fig. 1. The bilingual terminology extraction in Sketch Engine. The default sort criterion is logDice but the list of candidates can be sorted also by the co-occurrence frequency. The frequency numbers are links to monolingual concordances.

5 Gold standard and experiments

First, we tried to get existing gold standard data. We asked authors of two papers [7,11] for their gold standard data used for the evaluation but they

[9] `www.sketchengine.co.uk/wp-content/uploads/ske-stat.pdf`

could not provide us with it due to possible copyright issues. Another option was to use an official terminology base from the European Union institution Directorate-General for Translation IATE[10] (Interactive Terminology for Europe) and run the extraction on a translation memory from the same institution, DGT translation memory [1]. This is also not possible as the IATE is not fully compatible with DGT translation memory.

That is why we have prepared our own gold standard for English-Czech pair. We manually cleaned 1,000 term pairs from a run of our algorithm, optimized for a high-coverage output. We did not take into account whether the items in the list were actually terms or not as we didn't want to evaluate the monolingual terminology extraction. We only decided whether the terms were translated correctly and whether they covered the same scope of the term. This means that for example the translation of the term "dry linen content" – "obsah suchého prádla" was considered as correct while the translation "suchého prádla" of "dry linen" was considered incorrect. This particular error was caused by the monolingual terminology extraction step as it is an incorrect gender-dependent base form. The correct form "suché prádlo" was not found in this case. We removed 66% of the 1,000 term pairs.

The annotated terms were then divided into two files—a file with correctly translated terms (the gold standard) containing 328 term pairs and a file with incorrectly translated terms. The gold standard file then used to evaluate the output of different settings of the algorithm. We used standard metrics precision, recall, and F-1 score. For ABTE, recall is usually more important than precision as users expect to post-edit and clean up candidate lists but do not want to miss possible term candidates. Therefore, the modifications we used were designed to increase precision and at the same time not decreasing recall.

The modifications were: 1) preferred terms with fewer words in the first language L1, 2) different ratios of the number of characters in L1 and L2 and 3) discarding the terms with low co-occurrence.

The terms with fewer words were preferred by the following method. The formula uses two variables: the number of words in L1 and a coefficient by how much the shorter terms should be preferred. Smaller coefficient prefers shorter terms.

$$\frac{4 - \text{number of words}}{10 * \text{coefficient}} + 1$$

In the process of sorting the final list we multiplied logDice score numerator by the formula above. The second modification uses one variable ratio, which we call *ideal ratio*, and which states the ideal ratio between the length of words or phrases in L1 and L2. For example, the ideal ratio 0.8 means that the pairs in L1 which have 20% fewer characters than in L2 will be preferred. First, the real ratio, i.e. the ratio between the number of characters in the L1 and L2 is counted.

[10] `iate.europa.eu`

$$\text{real ratio} = \frac{\text{characters in L1}}{\text{characters in the L2}}$$

If the real ratio is smaller than the ideal ratio, we count the mod:

$$\text{mod} = \frac{\text{real ratio}}{\text{ideal ratio}}$$

If the real ratio is bigger than ideal ratio or equal, we count the mod:

$$\text{mod} = \frac{\text{ideal ratio}}{\text{real ratio}}$$

We then multiply logDice with the mod.

In the last modification, we discarded pairs with co-occurrence lower than 4. The gold standard does not contain any terms with co-occurrence under 4. Therefore, we did not decrease recall.

The gold data set is available for download[11] under CC BY-SA licence[12].

6 Evaluation

We tested the algorithm on the gold standard. While we tried different ratios, we found out that ratios 0.92–0.97 give the best results. This means that English terms should be slightly shorter than the Czech terms. It corresponds with the fact that English texts are approximately 1.1 times shorter than Czech.[13]

We also tested the preference of the terms having fewer words in the first language. The range of the coefficient which we tested was 1–6. However, all the settings where the coefficient was higher than 1 (the terms with fewer words had less preference) were significantly less successful. Furthermore, we tried to discard and to not discard the terms with low co-occurrence frequency. The results were better when the terms with low co-occurrence frequency were discarded. This was expected as mentioned above. All the settings discard the terms with low co-occurrence because those settings always yielded better results. The ratio column contains the ideal ratio. The first line contains original settings where the ratio and the shorter terms are not preferred.

The Table 1 shows that the best results were achieved with ratio between 0.92–0.97. We obtained 1.24 times better results with the experiments against the original settings.

7 Conclusions

We described an approach to ABTE already implemented in corpus management tool Sketch Engine. We also evaluated the algorithm using a manually

[11] `www.sketchengine.co.uk/bilingual-terminology-extraction`

[12] `creativecommons.org/licenses/by-sa/2.0`

[13] `www.eurotranslation.cz/index.php?menu=faq`

Table 1. Evaluation, the best results.

RATIO	PRECISION	RECALL	F-SCORE
N/A	0.3282	0.3232	0.3257
2.00	0.3994	0.3933	0.3963
0.98–1.5	0.4025	0.3964	0.3994
0.92–0.97	0.4056	0.3994	**0.4025**
0.91	0.4025	0.3964	0.3994
0.90	0.3994	0.3933	0.3963
0.80	0.3993	0.3933	0.3963
0.75	0.3963	0.3902	0.3932

prepared gold standard for English-Czech language pair from DGT translation memory and suggested a possible improvement of the method.

In the future we would like to try an alternative approach to ABTE in the form of an on-the-fly searching for translation candidates from parallel concordances for arbitrary phrases during a computer-assisted translation process.

The ABTE methods in general are not supposed to completely eliminate the need of a painstaking manual post-editing of terminology bases. The candidate lists are intended as a starting point for building bilingual terminology bases from scratch which can significantly speed up the process of otherwise tedious and time-consuming work.

Acknowledgement This work has been partly supported by the Masaryk University within the project *Čeština v jednotě synchronie a diachronie – 2015* (MUNI/A/1165/2014).

References

1. Steinberger, R., Andreas, E., Szymon, K., Spyridon, P., Patrick, S.: Dgt-tm: A freely available translation memory in 22 languages. Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012) (2012)
2. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The sketch engine: ten years on. Lexicography **1**(1) (2014) 7–36
3. Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P., Suchomel, V.: Finding terms in corpora for many languages with the sketch engine. EACL 2014 (2014) 53
4. Rychlý, P.: A lexicographer-friendly association score. Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN (2008) 6–9
5. Kupiec, J.: An algorithm for finding noun phrase correspondences in bilingual corpora. In: Proceedings of the 31st annual meeting on Association for Computational Linguistics, Association for Computational Linguistics (1993) 17–22
6. Kockaert, H.J., Steurs, F.: Handbook of Terminology. Volume 1. John Benjamins Publishing Company (2015)
7. Macken, L., Lefever, E., Hoste, V.: Texsis: Bilingual terminology extraction from parallel corpora using chunk-based alignment. Terminology **19**(1) (2013) 1–30

8. Itagaki, M., Aikawa, T., He, X.: Automatic validation of terminology translation consistency with statistical method. Proceedings of MT summit XI (2007) 269–274
9. Tiedemann, J.: News from opus-a collection of multilingual parallel corpora with tools and interfaces. In: Recent advances in natural language processing. Volume 5. (2009) 237–248
10. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT summit. Volume 5., Citeseer (2005) 79–86
11. Rösiger, I., Schäfer, J., George, T., Tannert, S., Heid, U., Dorna, M.: Extracting terms and their relations from german texts: Nlp tools for the preparation of raw material for specialized e-dictionaries. (2015)