Automatic Term Extraction from Turkish to English Medical Corpus

Gökhan Doğru[1]

Universitat Autònoma de Barcelona
08193, Barcelona, Spain
gokhan.dogru@uab.cat

Abstract. This study aims to evaluate the state-of-the-art automatic and semi-automatic term extraction from a domain-specific bilingual corpus in Turkish - English language pair. Three different tools (a computer-assisted translation tool, a web-based corpus analysis toolkit, and a desktop corpus analysis tool) are used for extracting Turkish cardiology single-word and multi-word candidate terms with different parameters, and the results are compared in terms of number of candidate term counts. It has been observed that each tool responds to different needs of translators and comes with a limited number of customization options. It is concluded that while monolingual term extraction is useful for translators and terminologists, there is still no tool providing bilingual candidate term extraction for Turkish – English language pair.

1 Introduction

Effective terminology management is one of the pillars of accurate, fast and consistent translation, a fact illustrated by the inclusion of terminology as one of the core categories of the widely used Multidimensional Quality Metrics (MQM) for translation quality assessment. Among other things, with the ubiquitous use of computer assisted translation tools and the necessity to have more than one translator in the workflow of translation require translation companies and other translation stakeholders to plan the terminology to be used across the projects. This planning phase is crucial to avoid any future inconsistency or inaccuracy in the use of professional terminology of a specific domain. Depending on the availability of previously translated documents (in the form of parallel corpora or comparable corpora), different methods of terminology management strategies including manual and automatic term extraction are implemented. In this study we report our results regarding automatic (candidate) term extraction from a Turkish- English bilingual cardiology corpus, the tools and methods used to process the term candidates and select actual terms, and the problems we have encountered.

EUROPHRAS 2019, pages 157–166,
Malaga, Spain, September 25-27, 2019.
https://doi.org/10.26615/978-2-9701095-6-3_021

2 Terminology Management and Automatic Term Extraction

Specialized domain translators may participate in ongoing translation projects in different time frames. Namely, although they may participate in the beginning of a project, they may also join in the second or third year to an ongoing project, e.g. a Spanish to Turkish medical device user manual translation with content updates regularly. In such a case, they should be provided by bilingual terminology lists (or glossary databases) prepared by terminologists or previous translators to be used within the preferred CAT tool. Especially, if many translators are to collaborate within the same project, these lists are particularly important to provide consistency. When such lists are not provided, translators need to prepare them by themselves as fast as possible to make their translation flow more efficient or at least, the translation company shall prepare it for them (preferably, by a terminology expert). But as Heylen and Hertog[1] observes, "(...) in a rapidly changing world with an ever-growing technical vocabulary, the manual maintenance, or in the case of new technological fields, the manual exploration, indexation and description of a domain's core vocabulary is a labor-intensive enterprise." Hence, bilingual and monolingual automatic term extraction methods have been suggested to make this preparation phase more agile and cost effective. These methods may be statistical, linguistic or hybrid depending on the tools used. We can also assume that more and more machine learning methods will be used for this process, yet it is possible to qualify machine learning methods under the category of statistical methods. Nevertheless, it should be stated that the results of the automatic extraction methods (just like the output of machine translation) are far from being perfect and as Ahmad and Rogers[2] emphasizes, "Term extraction produces the raw material for terminology databases: this raw material has to be examined, tested and validated in some way before inclusion in a terminology database." For this reason, the resulting lists of terms are called "candidate terms", not "terms" per se. Therefore, the complete process of term extraction is still a semi-automatic process, which needs human intervention. Although nearly two decades have passed since the article by Ahmad and Rogers[2] and new milestones have been reached in terminology extraction in different language pairs, this observation for automatic term extraction is still valid. In this time framework, terminology databases have become an inseparable component of computer assisted translation tools and more and more CAT tools are including automatic term extraction feature to their software. Besides, new use cases have been introduced for terminology databases including use in automatic translation quality assurance as well as machine translation training.

3 Methods and Tools

There are different software tools that allow for automatic candidate term extraction with or without the possibility to later edit or validate the candidates

in-situ. We have used three different tools for term extraction based on statistical frequency: MemoQ, a proprietary CAT Tool with term extraction and editing feature; AntConc, a stand-alone freeware corpus analysis toolkit; and Sketch Engine, a web-based corpus query and management system. We have used these tools to process a corpus consisting of cardiology journal abstracts in Turkish and their translations into English. The results of each tool are reported in the following sections. Among these tools, MemoQ has been used in the corpus building phase as well since it also includes features both for text parsing, translation memory and terminology management features.

4 Bilingual Corpus Use for Term Extraction

The rise of automatic term extraction is highly tied to the developments in electronic corpus studies. The tools and procedures used in monolingual and bilingual corpora preparation have made it easier to prepare domain specific corpora from which terminology can be extracted with certain levels of success. In translation world, the most common type of corpora are translation memories which include source language strings and target language strings together with some metadata including, date, domain, translator etc. These translation memories are saved and interchanged across different CAT tools in translation memory exchange format (TMX). Hence, in order to decrease the requirement for translators to use different stand-alone tools to realize term extraction tasks and avoid different file format exchanges (import/export) between terminology tools and translation tools, there is now a tendency to integrate term extraction capabilities into CAT platforms. One good example is MemoQ, a desktop CAT tool which integrates an automatic term extraction feature into the workflow of the translator within a project. In our study, we have a Turkish to English cardiology corpus prepared in the form of a translation memory. Using this corpus, we examine automatic term extraction tools and methods and obtain Turkish to English cardiology terms to be used in statistical machine translation. However, as we have mentioned above, these terms can also be used in translation projects to make translations faster, more consistent and of higher quality.

5 Term Extraction from Turkish to English Medical Corpora

Translating terminology properly and consistently is very vital in medical texts. Hence, the availability of bilingual terminological database (shortly, "termbase") is crucial in translation projects. Before describing our corpus, we have also investigated how we can ensure the quality or credibility of a termbase. Reynolds[3], in a professional translation context, divides terminology credibility into three: "a). High: "i). Terminology which the customer has specified should be used, ii). Terminology received from customer or acknowledged experts in the field terminology provided by the customer; b). Medium: i). Terminology verified by other

professional translators., ii). The more translators which agree that a particular term is correct, the more likely it is to be correct; Low: i). Terminology extracted using software tool, ii). Terminology received from other sources but not verified by the customer, industry experts or other translators". Below it will be seen that the corpora that we have prepared is derived from a cardiology journal with abstracts in Turkish translated to English. Since these abstracts and their translations pass from a peer review (including experts of the same domain) before publishing, we assume that the corpus terminology has a high terminology credibility according to the classification of Reynolds[3]. Although we use software for extraction of candidate terms, we also validate them manually.

5.1 Corpus Preparation

We have used different tools and methods to create a domain specific bilingual corpus. We build it from the abstracts published in *Archives of the Turkish Society of Cardiology* which "is a peer-reviewed journal that covers all aspects of cardiovascular medicine. The journal publishes original clinical and experimental research articles, case reports, reviews and interesting images pertinent to cardiovascular diseases, as well as editorial comments, letters to the editor, news, guidelines, and abstracts presented at national cardiology meetings."[1] Its topics include "coronary artery disease, valve diseases, arrhythmia's, heart failure, hypertension, congenital heart diseases, cardiovascular surgery, basic science and imaging techniques."[2] The journal's online archive dates back to 1990 and the current issue is Volume: 47 Issue: 3 - April 2019. Most importantly for our purpose, nearly all the abstracts are translated into English. Considering that the journal keeps being published for nearly three decades and that it includes scientific articles from cardiology domain, we can argue that it covers a significant portion of the Turkish cardiology terminology and, through the translations of the abstracts, the English counterparts of the terms as perceived by Turkish cardiology specialists. One of our purposes has been to truly represent the termbase available in this journal archive about the abstracts and their translations. The abstracts have a consistent format. Original articles include subheadings of "objective", "method", "results", "conclusions" and "keywords" (which is a valuable source for term extraction) while case report abstracts are shorter, have a keyword section and do not include subheading. Since the website has a well-structured HTML design, it has been possible to crawl all the abstracts (both in Turkish and English) in HTML format and save them locally for further processing. Then, we were able to convert the HTML files into plain text files using MemoQ's regex text filter (one custom filter for Turkish and a custom filter for English), and align them to build a translation memory.

[1] *Archives of the Turkish Society of Cardiology website*: http://www.archivestsc.com/about-the-journal (last access: 29.04.2019)¿

[2] Ibid.

The Turkish English Corpus in Numbers. The characteristics of our corpus are given in Table 1. Our corpus has 474,273 source language (SL) words and 542,783 target language (TL) words (Since Turkish is an agglutinative language with lots of suffixes added at the end of words, there is a 14,4 percent increase in the number of words in English). Ahmad and Rogers[2] suggest having a corpus of nearly 100,000 words "as a good starting point for corpus-based terminology management in a highly-specialized discipline" (p. 593) and later add that "As a rule of thumb, special-language corpora already start to become useful for key terms of the domain in the tens of thousands of words, rather than the millions of words required for general-language lexicography" (p. 594). Having 474,273 SL words and 57,368 unique SL words, our corpus seems to be satisfying the size criteria for terminology-oriented corpus. When we look at unique SL and TL word forms, we see a pattern similar to the total SL and TL word counts. While there are 57368 unique word forms in SL, there are 35.844 word-forms in TL[3]. For comparing how different corpus analysis tools show counts, we have also made a comparison of frequency and word counts in AntConc and Sketch Engine since how a word is defined can be different across tools and default tool settings.

Table 1. The profile of the corpus according to MemoQ.

Language pair	Turkish - English
Domain (field)	Medicine
Discipline	Internal medicine
Subdiscipline	Cardiology
UNESCO code	3205.01
Number of source words	474.273
Number of target words	542.783
Number of unique source words	57.368
Number of unique target words	35.844

A comparison of word and frequency counts is given below. It can be observed that each tool treats differently the concept of word; hence, each one yields different counts for both total word count and unique word count. Since results of one-word lists include around 40,000 words in each tool and there can be up to 5-word terms in this corpus, it is obvious that such a scenario is not terminologist-friendly and that it will be very time consuming to create a final terminology list. In the following sections, we will explain how we have constrained our term

[3] Unique word (form) counts are realized in Memoq's term extraction feature. Minimum frequency and maximum number of words per term are defined as "1" so that only unique words are extracted. For the initial analysis, any one or more letters are considered word forms. Of course, the term "term" is a different concept and it will be elaborated in the following sections.

Table 2. Comparison of word and frequency counts in 3 tools. *Word with number are ignored. **Non-words ("tokens which do not start with a letter of the alphabet.")[4]

	MemoQ	AntConc	Sketch Engine
No. of source words	474273	489155	479,200
No. of target words	542783	557054	523,077
No. of unique source words	40656	42836	38800
No. of unique target words	20802*	18309	23326**

extraction parameters in each tool to create a noise-free (or with minimum noise possible) candidate term list.

Constraining Parameters for Automatic Term Extraction. The genre of our corpus has a unique advantage: by nature, scientific abstracts include a section called "keywords" where the author(s) add(s) the most relevant keywords in their study. In each abstract, these keywords section is provided in a sentence.

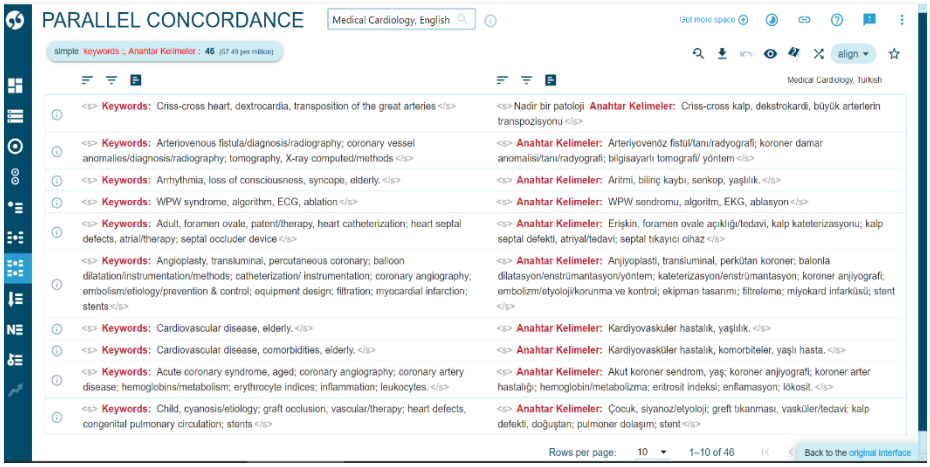

Fig. 1. Parallel Concordance in Sketch Engine.

In Sketch Engine, we have been able to filter these sentences and sort them side by side in Turkish and English version (called parallel concordance). A sample of this process is shown in the Figure 1. Then we could export all these

[4] `https://www.sketchengine.eu/my_keywords/non-word` (last access: 15.05.2019)

sentences to an excel file and then manually create a terminology list in cardiology domain. This bilingual list is very comprehensive because it includes most of the important terms inside the corpus. There are 2951 one-word and multi-word terms. And since these terms are not extracted but provided by cardiology experts, we can argue that they are reliable. We have used this manual list to compare the results of automatic extraction features of each tool.

6 Term Extraction from Turkish to English Medical Corpora

We have three different results for each tool. It should be stated that none of these tools allows for bilingual automatic extraction for Turkish and English language pair. Sketch Engine has this option for a few languages such as "Chinese, Czech, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish" as stated by Baisa[4] *et al.* but Turkish is not included. We believe that the addition of Turkish will make Turkish – English bilingual term extraction very fast and time-saving. In the sections below, we report our results for each tool. It will be observed that their different parameters and available features provide varying number of candidate one-word and multi-word candidate terms. It shall be emphasized that extracting more terms does not necessarily mean that the resulting list will be more useful for the translators since removing words that do no constitute terms can be time-consuming as well.

6.1 Term Extraction with MemoQ

MemoQ extracts multi-word and single-word terms monolingually based on frequency and confidence score and it is also possible to lookup the target terms through other active (reference) termbase lists. In our setting, we have the above-mentioned reference cardiology terminology to compare our results. Besides, we have realized that stop word lists are very important to avoid noise in term candidates and that MemoQ does not have a stop word list for Turkish. Stop word lists helps to filter out the words or groups of words that are frequent but are not actually terms. Hence, based on the initial candidate term results we have created a domain-specific stop word list. In other words, our stop word lists can be reused when extracting terms from medical corpora. We have made it openly available[5].

Setting 1: Multiword terms: Maximum 5 words, minimum 5 times frequency, Single-word terms: Minimum character length 4, minimum 5 times frequency; words with number ignored, term lookup is not active; stop word list is not used.

Setup 2: Multiword terms: Maximum 5 words, minimum 5 times frequency, Single-word terms: Minimum character length 4, minimum 5 times frequency;

[5] Stop Word List for Turkish language, `https://github.com/gokhandogru/stopwordsforturkish` (last access: 10/06/2019)

Table 3. MemoQ Candidate Term Extraction with Different Parameters.

Type of Extraction	Candidate Term Count	Full Match Term Lookup Count	Partial Match Term Lookup Count
Without Stop Word List	21557	334	5401
With a Stop Word List	13219	317	3631
Only Multi-Word Terms	5901	41	3088

words with number ignored, term lookup is active; stop word list is crafted and used.

Setup 3: Multiword terms: Maximum 5 words, minimum 5 times frequency, Single-word terms are ignored; words with number ignored, term lookup is active; stop word list is crafted and used.

6.2 Term Extraction with AntConc

AntConc allows for using a general monolingual corpus to compare with a domain specific corpus so that word(s) that occur more frequently in domain specific corpus and less frequently in the domain general corpus can be calculated as keywords or terms. We have used Turkish Wikipedia Corpus in OPUS Corpus as a reference corpus. This corpus has 4.7 million tokens. However, it has only been possible to extract single-word terms ("keywords") in this setting. We have yielded 5701 (candidate) keywords. For multi-word terms, we have tried the "Clusters/N-gram" feature with a setting of minimum frequency of 5 for terms with the size of 2 – 5 words. The system has yielded 20821 multi-word terms. However, this setting does not include a comparison with the reference corpus. Hence, it is only frequency-based and very noisy.

6.3 Term Extraction with Sketch Engine

Sketch Engine differentiates keywords from terms. Keywords are "individual words (tokens) which appear more frequently in the focus corpus than in the reference corpus" while terms are "multi-word expressions which appear more frequently in the focus corpus than in the reference corpus and, additionally, match the typical format of terminology in the language." Although this distinction is not that clear in Translation Studies and in Translation Technologies Studies, we will make our analysis with this distinction in mind. In the keyword extraction, Sketch Engine has yielded 7862 keywords. The reference corpus that they use is called Turkish Web 2012 and it includes more than 3 billion words. And the keywords are highly precise. As of June 2019, the Turkish multi-word term extraction has not been possible in Sketch Engine.

7 Results

Each tool has yielded a similar number of single-word term while the multi-word counts have differed from 0 to 20821 depending on the available configurations. On average, we have obtained 6960 one-word terms per tool. When it comes to Turkish multi-word terms, each tool has behaved differently. Firstly, Sketch Engine does not support multi-word term extraction for Turkish. We have experimented with the English corpus and the results have been very adequate. It will be very useful to have Turkish multi-word extraction option as well. While it is possible to extract multi-word terms with "Clusters/N-gram" feature, the results are too noisy because of the lack of a comparison with a reference corpus or linguistic normalization / filtering of the resulting noisy candidate multi-word terms. Excessive number of candidate terms is not going to be useful compared to manual term extraction. Lastly, Memoq has yielded a balanced number of multi-word terms after we have crafted a custom stop word list. The results given in the Memoq row in Table 4 reflects the extraction after the use of our stop word list. In all three tools, some of the Turkish terms are inflected, in other words, they include inflectional suffixes which shall be removed manually to lemmatize the terms before inclusion in the final term database. This remains as a problem to be solved in each tool. A lemmatization step in the automatic extraction stage can be a solution. The biggest productivity gain for translator can be achieved when bilingual candidate term extraction becomes possible. For now, none of these tools allow production-level Turkish-English or English-Turkish bilingual term extraction. Considering the increasing amount of translation between these two languages and the increasing need of collaborative translation, new techniques shall be developed to extract bilingual candidate terms.

Table 4. Monolingual one-word and multi-word term extraction in three tools.

	One-word terms	Multi-word terms
MemoQ	7318	5901
AntConc	5701	20821
Sketch Engine	7862	0

8 Conclusion

Turkish is a morphologically rich language with lots of suffixes in the end of the word roots, which has resulted in lots of noise in term extraction in all three tools that we have used since they all use conventional statistical methods. We think that the use of stop word lists for Turkish, growing use of deep machine learning (including neural networks) methods in term extraction as well as better strategies of lemmatization of Turkish words and multi-word units are going

to lead to better monolingual Turkish term extraction, which, in turn, will make bilingual term extraction possible for language pairs including Turkish. The new tools and techniques in natural language processing including neural machine translation provide an opportunity for bilingual term extraction and their integration into CAT tools can give translators an optimum work environment in terms of terminology.

Acknowledgements

This work has been funded and supported by the grant of AGAUR FI-2017.

References

1. Heylen, K., De Hertog, D.: Automatic Term Extraction. In: Kockaert, H. J., Steurs, F. (eds.) Handbook of Terminology, Vol. 1, pp. 203-221. John Benjamins Publishing Company, Amsterdam (2015).
2. Ahmad, K., Rogers, M.: Corpus Linguistics and Terminology Extraction. In: Wright, S. E., Budin, G. (eds.) Handbook of Terminology Management, pp. 725-760. John Benjamins Publishing Company, Amsterdam (2001).
3. Reynolds, P.: Machine translation, translation memory and terminology management. In: Kockaert, H. J., Steurs, F. (eds.) Handbook of Terminology, Vol.1, pp. 276-287. John Benjamins Publishing Company, Amsterdam (2015).
4. Baisa, V., Ulipová, B., Cukr, M.: Bilingual Terminology Extraction in Sketch Engine. In: Horák, A., Rychlý, P., Rambousek, A. (eds.) Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2015, pp. 61-67. (2015)
5. Multidimensional Quality Metrics (MQM). `http://www.qt21.eu/mqm-definition/definition-2015-06-16.htmlquality-terms`, last accessed 2019/04/25.
6. MemoQ Homepage, `https://www.memoq.com,lastaccessed2019/04/25.`.
7. AntConc Homepage, `https://www.laurenceanthony.net/software/antconc/`, last accessed 2019/04/25.
8. Sketch Engine Homepage, `https://www.sketchengine.eu/`, last accessed 2019/04/25.
9. Archives of the Turkish Society of Cardiology Homepage:`http://www.archivestsc.com/about-the-journal`, last accessed 2019/04/25
10. Opus Corpus Homepage, `http://http://opus.nlpl.eu/`, last accessed 2019/04/25