TERMINOLOGY EXTRACTION TOOLS: ARE THEY USEFUL FOR TRANSLATORS?

Hernani Costa
Anna Zaretskaya
Gloria Corpas Pastor
Miriam Seghiri

Terminology extraction tools have become an indispensable resource in education, research and business. Today, users can find a great variety of terminology extraction tools of all kinds, and they all offer different features. Apart from many other areas, these tools are especially helpful in the professional translation setting. We do not know, however, if the existing tools have all the necessary features for this kind of work. In search for the answer, we make an overview of nine selected tools available on the market and find out if they provide the translators' most favorite features.

Terminology extraction tools and their areas of application

The purpose of terminology extraction tools (TET) is to help users build terminological resources in a (semi-)automatic way. The need for such resources comes mostly from the growing needs in information management and translation, which make it more and more necessary to have some automated assistance when performing terminology-related tasks. Companies, freelancers and professionals in various linguistic fields can resort to these tools to, for example build glossaries, thesauri and terminological dictionaries that they use directly in their work. Moreover, TE is embedded in a number of natural language processing and linguistic research tasks, such as automatic indexing, machine translation, information extraction, creation of ontologies and knowledge bases, and corpus analysis. Although they have such broad range of applications, these tools are often designed for one specific purpose, which consequently makes their usage challenging when employed in a different setting.

One of the most important areas where terminology extraction is extremely helpful is in the translation industry. Today, more and more language service providers (LSP) as well as freelance translators and interpreters understand the benefits of automatizing terminology tasks. It not only allows them to quickly identify the domain of the documents they are dealing with, but also to easily find words and phrases that need to be paid special attention to. While translating terminological units, in many cases it is necessary to consider the domain and look up the term equivalents in special resources like terminology databases. And in addition, it helps maintain terminological consistency throughout the project between all the parts involved: the translator, the LSP and the client.

Apart from saving time, another significant advantage of using TET instead of manual terminology search consists in the possibility to specify different search criteria, which allows to adapt the search query to a particular task. This allows users to see all kinds of information they need about the term, and also to narrow the search and filter the results depending on what they are looking for. As an example, many state-of-the-art TET offer a possibility to see linguistic and statistic information about the term, the context where it appears, specify the number of words in the term, and many other useful features. Unfortunately, not every TET offers a full set of desirable features and settings, which makes it sometimes challenging to find the perfect tool for the task in hand. Apart from the functionalities they offer, TET also differ as to the environment they work in. For instance, standalone installable tools require an installation process and work as independent computer programs. There also exist web-based tools, which work within a browser. And finally, there are reusable software that facilitates the development of larger applications, called frameworks.

Considering the existing variety, it is not clear how a professional translator is to proceed when choosing a TET suitable for the job. As we will see further, there are some TET that are specifically created for translators. But do they have all the necessary characteristics for translators? And, furthermore, what exactly are these characteristics?

Standalone Terminology Extraction Tools

Standalone software is probably the most popular type of software today, and TET are no exception. Standalone TET are tools that can be installed on the computer and operate independently of any other device or system.

SDL MultiTerm Extract is one of such applications. It is a component of SDL MultiTerm, a commercial terminology management tool that provides one solution to store, extract and manage multilingual terminology. Multiterm exists as a standalone application, and can also be integrated in SDL Trados Studio. It is one of the few tools that were designed specifically to be used by translators and is probably the most well-known TET in the translation industry. This TE system locates potential monolingual and bilingual terminology in documents and translation memories using a statistic-based method. The user can validate the extracted candidate terms by looking at a monolingual or bilingual concordance. A big advantage of this tool is its support for any language, including Unicode languages. In addition, it offers a number of functionalities that are useful in different translation scenarios, such as ability to compile a dictionary from parallel texts; flexible filtering that ensures that only the most frequent candidate terms are extracted; possibility to store an unlimited number of terms in any language; import and export glossaries from and to different technology environments. In addition, its integration with SDL Multiterm gives access to many convenient term-management functions, such as manually adding a variety of meta-data information to the terms, such as synonyms, context, definitions, illustrations, part-of-speech tags, URLs, etc., and searching not only the indexed terms but also their descriptive fields.

Simple Extractor, as its name implies, offers significantly less functionalities compared to the previous tool. It is a commercial TET developed by DAIL Software S.L. for Mac OS, Linux and Windows platforms. This clean and easy-to-use standalone Java application was designed to automatically extract the most frequent words and multi-word terms from English, Portuguese, Spanish, French and Russian documents. Simple Extractor not only permits to extract a list of terms (from unigrams up to seven-grams), but also specify the minimum and maximum number of occurrences of a term. Moreover, Simple Extractor offers an option to load stopword lists, an advanced search functionality that permits to search through the extracted list of terms, to explore all the contexts that a specific term appears, to edit the term text, to filter the extracted terms according to the number of words that form them, and to sort the displayed output by any of its fields (frequency, term and context in alphabetical order). Finally, Simple Extractor permits to print out or export to a file (.pdf, .doc, .csv or .txt) all the extract terms, as well as their frequencies and corresponding contexts.

TermSuite is an open-source and platform-independent TET written in Java and distributed under the Apache License 2.0. It was developed within the scope of the TTC (Terminology Extraction, Translation Tools and Comparable Corpora) project, whose purpose was to design a tool capable of extracting bilingual terminology from comparable corpora in seven languages: English, French, German, Spanish, Chinese and Russian. TermSuite's architecture is composed by 3-step modules: the Spotter, the Indexer and the Aligner. The Spotter module is responsible for preprocessing the input monolingual corpus, i.e., it performs tokenization, part-of-speech tagging, stemming and lemmatization. Then, the Indexer module uses both a statistic and a linguistic-based approach to extract monolingual terminology from a monolingual corpus processed by the Spotter. Finally, the Aligner computes the translation of a source terminology into a target language. The source and target terms required are these already computed by the Indexer module, which means that the previous two steps should be repeated for the target language. The user can choose from several alignment options, such as the selection of the maximum number of translation candidates for a given source term, the use of similarity measures to compare the contexts of the term in the source and the target languages, amongst other advanced settings. Once all the parameters are set, it is possible to view and explore all the translation candidates ranked according to their similarity score within the tool or use the output XML file for other purposes.

Web-Based Terminology Extraction Tools

Although standalone TET still are predominant on today's TE applications market, the future web-based TE technologies will certainly evolve by migrating all standalone features to a web-based environment, which will allow them to consequently take over the leadership in the near future. As we will see, there are already some examples of this trend. The advantages are that web-based TET, compared to standalone tools, do not require any prior installation as they can be accessed within a web browser and that they make use of web technologies. Although most of web-based TET are often integrated as features in cutting-edge web-based applications with a wider purpose, such as managing corpora or terminology (e.g., Sketch Engine and Terminus, respectively), there also exist tools like the TET by Translated, which were developed with the proper purpose of terminology extraction.

Sketch Engine is an online tool created by Lexical Computing Ltd for building and managing corpora, which along with a number of corpus-processing features includes terminology extraction. It can be accessed under a paid commercial or academic license and supports 82 languages. This tool offers both monolingual and multilingual extraction. When extracting monolingual terminology, the user can choose whether to extract only single words (keywords) or multi-word terminological units (terms). In the output, the user can see the keywords or terms, links to the five most relevant Wikipedia articles for each of them, the term's score, its frequency in the searched corpus, and its frequency in the reference corpus. There are a variety of search options that can be tuned. For instance, the user can choose a different reference corpus, decide whether search for words or lemmas, and accentuate low or high-frequency keywords according to the preferences. The output can be downloaded as a TBX or CSV file. In order to perform multilingual term extraction the user needs to upload a TMX file with a parallel corpus aligned on the sentence or paragraph level. The terminology is first extracted within each language resulting in lists of candidate terms. In the second step, the system searches for such pairs of candidates which co-locate in the parallel documents most often. The resulting list of candidate pairs (terms in two languages) is then presented to the user. Results can be saved in a TBX or TXT file, which is especially convenient for computer-assisted translation tool users.

Translated s.r.l., a leading LSP developed a web-based tool that can be accessed directly on the company's website. It was created in order to help translators with their translation jobs by identifying the difficulties in the text and simplifying the process of creating glossaries. Up to the current date it supports only English, Italian and French. The system output includes the top 20 terms ranked by their score. In addition, the terms are given as hyperlinks to the corresponding Google search results.

Below the list of terms the tool also shows all the terms in their full-sentence context. In order to easily differentiate the terms, each term is highlighted by a different color. In general, this tool is quite simple compared to the others, but can provide a fast and free solution any time it is needed.

Terminus is a web-based application for corpus and terminology management developed at the University Pompeu Fabra, Spain and it can be accessed by software licensing. The purpose of this tool is to integrate the complete process of terminographic work: textual corpus search, compilation and analysis, term extraction, glossary and project management, database creation and maintenance, and dictionary edition. This is done with the help of a number of articulated modules, including the Analysis module, which has a semi-automatic term extraction feature. The extraction process has two options: the user can train a term extractor in a specific domain by incorporating an electronic dictionary containing terms of the same field, or simply apply a generic ready-to-use term extractor to any textual corpus. In addition, one can use other features to extract term candidates, such as the n-gram extractor, bi-gram extraction with association measures, keywords, and later manually validate relevant terms.

Frameworks

Frameworks are different from the other two types of tools because they are not complete software products but reusable software environments or libraries that can be used or even completely integrated in larger translation software applications, products or solutions. In particular, systems of this type are often used in information retrieval, where identification and indexing of terminology serves as an aid to information retrieval queries. In detail, the purpose of terminology extraction for both information retrieval and document retrieval is to isolate terms that contain enough informational content to support retrieval based on the queries supplied when querying a set of documents.

Keyphrase Extraction Algorithm (Kea) is a framework specially designed for automatically assigning terms to a document (aka keyphrase indexing). Kea is a platform-independent toolkit implemented in Java and distributed under the GNU General Public License. In detail, this framework can either be used for free indexing or for indexing with a controlled vocabulary. When used as free indexing, Kea looks for significant terms in a document. If on one hand, the free indexing option can be applied to any document and working language (as long as the corresponding stopword file and stemmer are provided). The controlled indexing, on the other hand, has the advantage that all documents are indexed in a consistent way disregarding their wording as the algorithm only collects those n-grams that match thesaurus terms.

Rainbow is a simple, yet powerful open-source platform-independent terminology extraction tool written in Java that uses statistic-based methods to automatically extract terms from multiple files and formats in any language. It is based on the Okapi Framework, a free, open-source and cross-platform framework that has a set of components and applications designed to help engineers, developers, translators and project managers involved in localization and translation-related tasks.

Java Automatic Term Extraction (JATE) is a JAVA toolkit that comprises several state-of-the-art term extractions algorithms. The motivation of this TET is three-fold: make available several automatic term extraction algorithms for the research community; encourage developers to built their methods under a uniform framework; and, enable comparative studies between different term extraction algorithms. JATE's workflow follows the typical TET steps: extract candidate terms from a corpus using linguistic tools; extract the candidates statistical features from the corpus; and, apply automatic terminology extraction algorithms to score the candidate terms domain representativeness based on their statistical features. So far, JATE's current version includes twelve state-of-the-art statistical algorithms.

Translators' preferences and opinions on the features of TET

As we mentioned above, translation is one of the most important applications of terminology extraction. However, it has not yet become a common part of the professional translation workflow. This was demonstrated by a user survey replied by over 600 translation professionals (Zaretskaya et al., 2015), which showed that only 25% of the respondents regularly resorted to TE in their work. It could be due to unsatisfying performance of the existing tools, their interface design, or simply to translators’ lack of awareness of these tools and of the benefits they can yield.

We have already seen that TET can differ as to various characteristics, such as their interface type (standalone, web-based or reusable libraries), document formats they support, languages they work with, as well as different search options. According to the survey findings, 27% of the respondents preferred to have a TE feature within their computer-assisted translation (CAT) tool instead of a separate TE software. Some translators, however, preferred a web-based application (9%) or installing a standalone tool on their computer (8%). Nevertheless, the majority (56%) reported that they did not have any preference regarding the tool's interface. The fact that translators prefer to have a TE system integrated in their CAT tool is related to the general tendency of CAT tools to include more and more different features. Indeed, translators have to deal with a great number of tools that help them automatize different stages of the translation process, so they prefer having one tool with multiple functions rather than having to look for and in many cases pay for several tools.

Regarding the importance of the TE's features, the most useful feature according to survey's participants was bilingual term extraction. In fact, considering that within a translation workflow, terminology extraction is performed with the final objective to translate the extracted terms, it is more convenient to have the terms extracted in the two languages simultaneously. Bilingual extraction is much harder to perform than only monolingual as it requires a good word alignment system, so not many existing tools offer this feature. In particular, among the tools we considered in the previous section only SDL Multiterm Extract and Sketch Engine have bilingual extraction. Similarly, TermSuite also offers translation candidates for the extracted monolingual terms, which is a different procedure, but still leads to the same results: terms in two languages. The second ranked feature was the possibility to compare the context of the term in the source and the target language, which is another type of bilingual analysis suitable for the translation task. This feature is also quite rare, and of all the considered tools, only SDL Multiterm Extract allows such analysis. The possibility to validate terms or, in other words, choose the terms that should be extracted instead of extracting all terms was ranked third and is also considered useful for translators. This feature is offered by almost all systems, except for TermSuite and Translated. Compiling a bilingual dictionary from parallel texts is another useful feature, which is offered only by SDL Multiterm Extract and by TermSuite. Finally, the respondents considered it useful to extract context together with terms or to see examples from the corpus. This is a common feature for many of the studied tools, including SDL Multiterm Extract, Simple Extractor or Translated.

Other features that were considered included support for different file formats, possibility to sort terms by frequency, support for many languages, possibility to specify the minimal number of occurrences of the words, show linguistic information about the term, and select the maximum number of translations for one term. All of them were considered useful, but were not among the most useful features.

	SDL Multiterm	Simple Extractor	TermSuit	Sketch Engine	Translated	Terminus	Kea	Rainbow	JATE
Bilingual extraction	✓		✓	✓					
Source and target context comparison	✓								
Terms validation	✓	✓		✓		✓	✓	✓	✓
Bilingual dictionaries compilation	✓		✓						
Context extraction	✓	✓	✓		✓	✓	✓	✓	✓
Support various file formats	✓	✓	✓	✓		✓	✓	✓	✓
Rank terms by frequency	✓	✓	✓	✓		✓	✓	✓	✓
Support for many languages	✓		✓	✓		✓	✓	✓	✓
Specify the minimal number of occurrences	✓	✓		✓		✓	✓	✓	✓
Show linguistic information	✓		✓			✓			
Specify the maximum number of translations			✓						
Stopword list option	✓	✓			✓		✓	✓	✓
Choose the minimum and maximum number of words per term	✓	✓					✓	✓	✓
Term statistics	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison chart of features for the selected tools.

And finally, some features were not considered so important by the respondents. One of them was the stopword list option: some of the tools, like Simple Extractor, allow to choose whether to use a stopword list, and others use it by default. Choosing the minimum and the maximum number of words per term, which was also among the least useful features, can be tuned by all the mentioned TE frameworks, for example. And finally, term statistics, which to some extent are provided by all tools, were not very important for most translators either. Table 1 shows which of the aforementioned features are presented in the 9 selected tools.

	Availability	Notes
SDL Multiterm	**$ 500**	**Free demo available**
Simple Extractor	**$ 140**	**60-days demo**
TermSuite	**Open-source**	
Sketch Engine	**$ 65/ year**	**30-days demo**
Translated	**Free**	
Terminus	**$ 440/ year**	**15-days demo**
Kea	**Open-source**	
Rainbow	**Open-source**	
JATE	**Open-source**	

Table 2: Depending on the purpose the quotes may vary. This table only shows the prices for licenses paid by individuals.

Conclusion

Although terminology extraction plays an important role in several disciplines such as linguistic research or language teaching, it is in the field of translation, particularly in the translation industry where its advantages are fully exploited and integrated in the workflow. An example of that is the use of bilingual term extraction, compiling dictionaries and comparing context in different languages as essential features for translators' work. In addition, it is also very useful for translators to see the terms in their context in order to understand their meaning and be able to find an adequate translation equivalent. Not all existing tools, however, provide these functionalities. We suggest that developing TET more suitable for the purpose of translation could help professionals in the industry take better advantage of TE technology. This has to be done, first of all, by taking into account the user requirements. As a step further in this direction, it would be necessary to investigate in more detail translators' attitudes towards TE tools. Especially, the reasons that prevent the vast majority of professional translators to adopt them. For instance, many translators might not be aware of their existence or understand their purpose, do not have time to learn how to use another complicated interface, or simply have other established procedures for dealing with terminology.

Acknowledgements

Hernani Costa and Anna Zaretskaya contributed equally to this work and are both supported by the People Programme (Marie Curie Actions) of the European Union’s Framework Programme (FP7/2007-2013) under REA grant agreement n.317471.

Bibliography

Zaretskaya, Anna and Corpas Pastor, Gloria and Seghiri, Miriam. 2015. "Translators' requirements for translation technologies: a user survey". *New Horizons in Translation and Interpreting Studies (Full papers)*. Tradulex. Genebra, Switzerland. December, 2015. pp.247-254. ISBN:9782970073659.