September 1, 2012

10 Things you should know about automatic terminology extraction

Uwe Muegge

10 Things You Should Know About Automatic Terminology Extraction

By Uwe Muegge

It is probably safe to say that many, if not most, commercial translation and localization projects today are carried out without a comprehensive, project-specific, up-to-date glossary in place. I suspect that one of the primary reasons for this inefficient state of affairs is the fact that many participants involved in these projects are unfamiliar with the tools and processes that enable linguists to create monolingual and multilingual glossaries quickly and efficiently. Below are 10 insights for linguists wishing to give automatic terminology extraction a(nother) try. (Please see the links in the box on page 27 for information on all of the tools mentioned here.)

> As using free MT services is becoming more and more popular among professional translators, so is the desire to control terminology in the final output that is delivered to clients.

1. The two biggest issues with terminology extraction tools are noise and silence. Many commercial terminology extraction tools use a language-independent approach to terminology extraction, which has the benefit of giving linguists a single tool for extracting terminology in many different languages. The drawback of this approach is that the percentage of "noise" (i.e., invalid term candidates) and "silence" (i.e., missing legitimate term candidates) is typically higher than in linguistic extraction tools that use language-specific term formation patterns. As a result, many linguists who use these popular extraction products are disappointed by the amount of clean-up work that some of these fairly expensive products can require.

2. For short texts, manual extraction may be your best option. To the best of my knowledge, there is no automatic terminology extraction system—at least for the English language—that creates term lists reliably

without requiring substantial human intervention either prior to extraction (e.g., set-up, importing word lists, creating rules, etc.) or after extraction (primarily manual or semi-automatic clean-up). For this reason, short texts are typically not well suited for automatic terminology extraction. (What qualifies as "short" differs from tool to tool, but 1,000 words serves as a general guideline.) This rule holds particularly true when the person performing the term extraction will subsequently translate the source text. It is generally a good idea to read the text to be translated in its entirety before translation, which creates a perfect opportunity for manual terminology extraction.

3. Rule-based MT systems are a great choice for low-cost automatic terminology extraction. Rule-based machine translation (MT) systems are among my favorite translation tools. Unlike statistical MT systems, rule-based MT products do not require any linguistic training on bilingual data to be useful, but rely on built-in grammar rules for the analysis of the source and generation of the target. More than 10 years ago, at the translation quality conference TQ2000 in Leipzig, I presented a paper on how to use rule-based MT systems to perform automatic terminology extraction.[1] One would expect that after so many years, it would now be common knowledge that the "Unknown Word" feature of rule-based MT systems is highly suitable for automatic terminology extraction. But, unfortunately, it just is not so. So let me tell you again: If you are a freelance translator or small translation agency, the most powerful, customizable, and cost-effective terminology extraction solution you can buy is a rule-based MT system. My two recommendations for this category are Systran Business Translator (available for 15 languages; Price: US$299) and PROMT Professional (available for 5 languages; Price: US$265). Both of these translation tools are very mature. They also offer a built-in translation memory and very large general and subject-specific dictionaries that make these products a great investment for any professional translator working in a covered language combination.

> Many commercial terminology extraction tools use a language-independent approach to terminology extraction.

4. Some free translation memory systems offer excellent built-in automatic terminology extraction. Similis is an often overlooked, yet extremely capable, free translation memory system. Since Similis, much like a rule-based MT system, uses language-specific analysis technology, the quality of the term extraction lists that this translation memory product generates puts it in a class of its own among translation memory systems. One particularly useful feature of Similis is its ability to extract highly accurate bilingual glossaries from translation memory (TMX) files. If you work from English and a half-dozen other supported languages, this might be the terminology extraction tool for which you have been looking.

Another translation memory solution that is available at no cost to freelance translators and students is Across Personal Edition, which includes crossTerm, a full-featured terminology management module complete with a statistical terminology extraction function. Unlike Similis, the Across tools support a wide range of languages and language combinations.

5. Use a concordance tool for simple terminology extraction. Stand-alone concordance tools have been used as research tools in corpus linguistics for a long time. A concordancer is a type of software application that allows users to extract and display in context all occurrences of specific words or phrases in a body of text. While concordance software is typically used to study collocations, perform frequency analyses and the like, linguists can use, and have been using, concordancers for terminology extraction. One of the best concordancers for terminology extraction is AntConc. This tool is highly customizable. For example, it allows users to define the word length of terms and supports multiple platforms (i.e., Windows, Mac, and Linux). It is also free.

6. Free online tools provide powerful terminology extraction, and there is nothing to install. If you are still not convinced that automatic terminology extraction is for you, let me introduce you to a set of tools where all you have to do to create a term list is to specify a source text and then press a button or two. There is no software to install, no manual to read, and, of course, no price to pay. With web-based terminology extraction

services like TerMine and FiveFilters Term Extraction, automatic terminology extraction really is child's play.

Do not let the simple interface of these sites fool you. Both of these online tools produce professional quality extraction lists that include compound nouns, and, in the case of TerMine, even scored rankings of term candidates.

7. Are you using free MT? Start post-editing with a glossary. As using free MT services is becoming more and more popular among professional translators, so is the desire to control terminology in the final output that is delivered to clients. Google Translator Toolkit is a free, full-featured online translation memory system that allows users to post-edit translations generated by Google Translate, Google's proprietary MT system. Since Google Translate is a statistical MT system that has been, and continues to be, trained on a wide variety of documents, the same source term might get translated in multiple ways even within the same document, not to mention across documents.

While it is currently not possible to submit user glossaries to Google's MT engine, it is possible to upload glossaries to the Translator Toolkit. And using one of the tools mentioned in this article to extract terminology and build a bilingual glossary before translating/post-editing in Google Translator Toolkit may be the best thing linguists can do to improve the efficiency of an already very efficient process.

8. Clean up your terminology extraction list to identify the most important term types. In my professional experience, term lists generated by automatic terminology extraction tools are never perfect. Even the best term extraction systems introduce "noise." For example, in the TerMine term list shown in Figure 1 below, I would argue that at least three of the 10 term candidates in this list require editing.

While most illegitimate term candidates are easy to identify (e.g., misspelled, truncated, incorrectly hyphenated words), many linguists have a hard time answering the following question: Which term candidates should users of terminology extraction systems actually develop into multilingual glossaries? There is no simple solution to this problem, as each translation project has its own limiting factors, available time typically being the most important one.

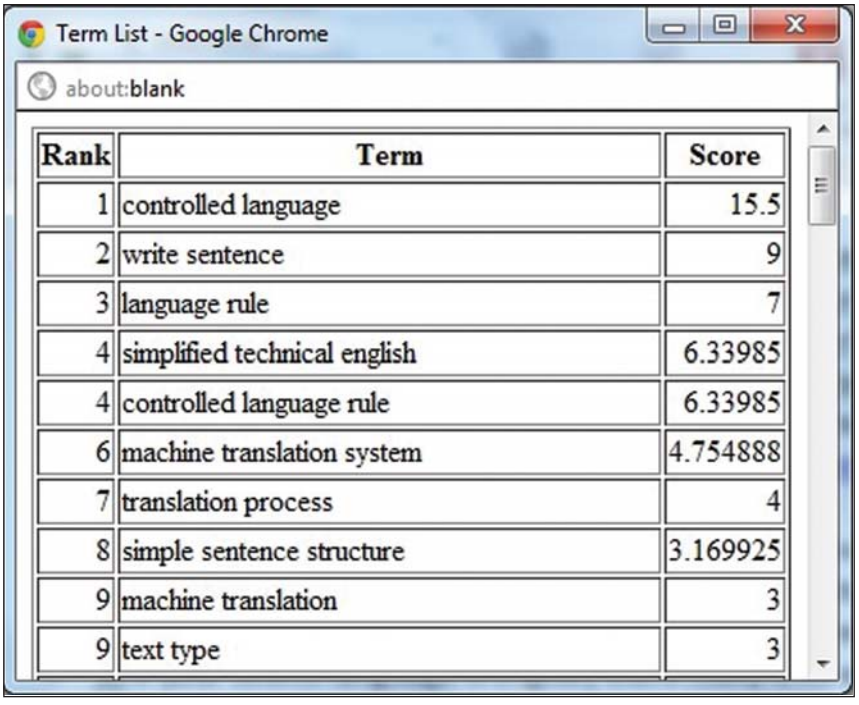
My recommendation for commercial translation projects is *always* to include the following types of terms in a project glossary, *even if the term occurs only once in a source document*. Mandatory term types include:

- Client business names;
- Product names; and
- Trademarks.

Yes, I know, this piece of advice runs counter to what many other terminology experts say; namely, if a term occurs only once in a text, there is no risk of inconsistency, and therefore single terms should not be included in glossaries. To that I say: There are terms that are so important that if a linguist gets them wrong, *even just once*, it would be a huge embarrassment for all parties involved.

Including every term of the above-mentioned types is particularly important when working with MT, as MT systems are notorious for "making-up" their own terminology in the target language. As such, the following term types should be included in glossaries based on the frequency

Figure 1: A sample term list generated by TerMine, a free online terminology extraction service.



Rank	Term	Score
1	controlled language	15.5
2	write sentence	9
3	language rule	7
4	simplified technical english	6.33985
4	controlled language rule	6.33985
6	machine translation system	4.754888
7	translation process	4
8	simple sentence structure	3.169925
9	machine translation	3
9	text type	3

of their occurrence in the source text (many terminology extraction tools provide frequency information):

- Feature names;
- Function names;
- Domain-specific terms; and
- Generic terms.

9. Use the recommended data categories when integrating an extraction list into a terminology management system. Once the extraction list has been cleaned up, the next logical step is to develop a multilingual glossary that will add value not only to the translation process but ideally to the entire translation cycle. The most valuable glossaries are those that provide information that goes beyond simple word pairs of "source term" and "target term." Here is the minimum data model I recommend for commercial projects:

- Client and/or business unit and/or project name;
- Source term;
- Part of speech (e.g., noun, proper noun, compound noun, verb, adjective, other);
- Context (e.g., a sample sentence in which the source term occurs); and
- Target term.

The big question at this stage is: What software platform do we use for developing and managing terminology after extraction? This is an important question as many, if not most, linguists do not have a proper terminology management system in place. While it may be tempting to use Microsoft Word or Excel tables to manage terminology—after all, these are programs that most linguists own and know how to use—word processors and spreadsheet applications are not good choices for managing ➡

Links Related to Tools

Systran Business Translator
http://owl.li/ciGG5

PROMT Professional
http://owl.li/ciGQI

Similis Free Download
http://owl.li/ciHfQ

Similis Terminology Extraction How-To Information
http://owl.li/ciHxp

Across Personal Edition
http://owl.li/ciHPf

crossTerm
http://owl.li/ciHZ5

AntCoc
http://owl.li/ciIvf

TerMine
http://owl.li/ciIH9

FiveFilters Term Extraction
http://owl.li/ciIQT

Google Translator Toolkit Registration Page
http://owl.li/ciJbr

TermWiki
http://owl.li/ciJr6

TermWiki Pro
http://owl.li/ciJyV

terminology data. The systems I recommend are TermWiki (if you are willing to share terminology) and TermWiki Pro (if you need to keep your terminology data private). Full disclosure: I have been, and keep, contributing to the development of TermWiki, which is already changing the way thousands of users around the globe manage linguistic assets.

Here are some of the benefits of using either version of TermWiki:

- Completely web-based (no software to install).
- Platform-independent (runs on Windows, Mac, Linux, Android, iOS, etc.).
- Wiki user interface (intuitively familiar, easy-to-use).
- Powerful collaboration features (automatic workflow management, etc.).
- No-cost/low-cost solution (TermWiki is free, TermWiki Pro is US$9.95/user/month).

10. A small investment in automatic terminology extraction can yield a big return in efficiency and client satisfaction. Being able to extract terminology quickly and efficiently is a wonderful thing. With automatic terminology extraction as part of a comprehensive terminology management effort, you are able to:

- Create comprehensive multilingual glossaries *before* translation.
- Have the client authorize project-specific, multilingual glossaries *before* translation.
- Have translation memory systems automatically suggest authorized translations for every term *during* translation.
- Have all members of a translation team use the same terminology *during* translation.
- Eliminate (terminology) review and corrections *after* translation.

> If you are a freelance translator or small translation agency, the most powerful, customizable, and cost-effective terminology extraction solution you can buy is a rule-based MT system.

With so many powerful terminology extraction tools to choose from, as long as the source language is a major language, there really is no excuse for *not* extracting terminology and creating a glossary as part of every translation project.

Notes

1. The material presented here was inspired by a two-part series I wrote for T for Translation, the blog of CSOFT International, blog.csoftintl.com. If you read German, an expanded version of this article is available at: http://owl.li/ciGr9.