Ten things you should know about automatic terminology extraction (Part I)

Uwe Muegge
Ten Things You Should Know about Automatic Terminology Extraction (Part One)

Welcome to Part One of a two-part series on automatic terminology extraction! Written by Uwe Muegge, CSOFT’s very own terminology management wizard, the five points in this post will supply readers with practical observations about and aspects of automatic terminology extraction. Don’t forget to stay tuned for next week’s Part Two, which will suggest different tools and methods to better manage terminology extraction.

Part 1 of 2

10 THINGS YOU SHOULD KNOW ABOUT
Automatic Terminology Extraction

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 not familiar with the tools and processes that enable linguists to create monolingual and multilingual glossaries quickly and efficiently. Below are five valuable insights for linguists who wish to give automatic terminology extraction another try.

1. The two biggest issues with terminology extraction tools: Noise and silence

Many commercial terminology extraction tools, including SDL MultiTerm Extract, 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 reliably creates term lists 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 1000 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 machine translation systems are a great choice for low-cost automatic terminology extraction

More than ten years ago, at the translation quality conference TQ2000 in Leipzig, I presented a paper on how to use rule-based machine translation systems to perform automatic terminology extraction (if you read German, here is a link to an updated version of this paper). One would expect that, after so many years, it is now common knowledge that the ‘Unknown Word’ feature of a rule-based MT system is highly suitable for automatic terminology extraction. Unfortunately, this just isn’t the case.

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 machine translation system. My two recommendations for this category are Systran Business Translator (available for 15 languages, costs approximately US$299) and PROMT Professional (available for 5 languages, costs approximately US$265).

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 machine translation system, uses language-specific analysis technology, the quality of the term extraction lists generated by this TM product places it in a class of its own among translation memory products. 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 you have been looking for. You can download Similis here.

Another translation memory solution that’s 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 have been using concordancers for terminology extraction. One of the best concordancers for terminology extraction is AntConc. This tool is highly customizable, e.g. it allows users to define the word length of terms, supports multiple platforms, i.e. Windows, Mac, and Linux, and is available for free.

Stay tuned for Part Two.

Uwe Muegge has more than 15 years of experience in the translation and localization industry, having worked in leadership functions on both the vendor and buyer side. He has published numerous articles on translation tools and processes, and taught computer-assisted translation and terminology management courses at the college level in both the United States and Europe. Uwe has been with CSOFT since 2008, and currently serves as Senior Translation Tools Strategist for North America.