Large-Scale Bilingual Extraction and Validation of Structured Patent Terminology

Magnus Merkel
Linköping University / Fodina Language Technology AB

A leading IT company in the world expresses its goals as making all information available to everybody, anywhere and anytime. One of the obstacles to achieving this goal has to do with language and language barriers. Automated translation (or machine translation MT) has been a field of study since the fifties within AI and has been seen as the holy grail of language technology for almost as long. MT is a key component for the aforementioned company if they are to succeed in bringing information across language barriers.

In reality, there are two major directions in MT today. One is data-driven and focused on statistical processing of documents. The other is more traditional, and based on rule-based translation systems, where linguistic knowledge is encoded in large lexicons and grammar rules. The data-driven camp is convinced that more data will solve the problems, e.g. by feeding a statistical MT system with tens of millions of parallel sentences (original and the corresponding translations) the statistical machinery will be able to create language and translation models that will produce high-quality translations. The rule-based MT camp believes that there are inherent features of human language that can never be modelled by massive amounts of data, and furthermore, that there simply are not enough parallel data for many language pairs to be found.

One subject field where high-quality automatic translations would be extremely useful concerns the patent area. Patent information is crucial for many businesses and to obtain patents and protect products are costly, for many reasons. The European Patent Office (EPO) has launched an automatic translation service on the Internet where patent agents can search approved patent applications and have them translated into several languages. This service is intended to be expanded to cover all European languages at the end of the project.

In this talk I will describe a large-scale extraction project of patent terminology (English-Swedish), which will be plugged into the EPO web service during the summer of 2009. The MT architecture on which the service is built is a rule-based architecture, where English is used as a pivot language and modules for translating

between any language X and English is being built. Starting in the summer of 2009, patent terminology was extracted from a set of English-Swedish document pairs. Information on the correct linguistic inflection patterns and hierarchical partitioning of terms based on their use was of utmost importance.

The process contains six phases, 1) Automatic analysis of the source material and system configuration; 2) Automatic term candidate extraction; 3) Term candidate filtering and initial linguistic validation; 4) Manual validation by domain experts; 5) Final linguistic validation; and 6) Publishing the validated terms.

Input to the extraction process consisted of more than 91.000 patent document pairs in English and Swedish, 565 million words in English and 450 million words in Swedish. The English documents were supplied in EBD SGML format and the Swedish documents were supplied in OCR processed scans of patent documents. After grammatical and statistical analysis, the documents were word aligned. Using the word-aligned material, candidate terms were extracted based on linguistic patterns. 750,000 term candidates were extracted and stored in a relational database. The term candidates were processed in 8 months resulting in 181.000 unique validated term pairs, which were then exported into several hierarchically organized terminology files, to be plugged into the rule-based MT system.