Evaluation of Automatic Term Alignment

There are numerous approaches for the identification (i.e. alignment) of equivalent domain-specific terms across languages (e.g. (DE) Strom ↔ (EN) electricity). Typically, the alignment provides a list of ordered target language terms for each source language term. In the literature, the performance of the alignment is measured in many different ways which makes a comparison of the approaches quite difficult (Laroche and Langlais, 2010). Usually, the evaluation makes use of a gold standard, i.e. manually established list of term equivalences which is then compared with the output of the alignment tools.

In the project TTC (Terminology extraction, Translation tools and Comparable corpora), we developed and tested different alignment methods using a common set of comparable corpora for the computation of the alignment, as well as a gold standard. The performance of the alignment tools was tested on several different language pairs. Even such a controlled evaluation has proven not to be unproblematic:

1. The corpora used for the alignment have to match the gold standard: the terms in the gold standard must all occur in the corpora used by the tools under evaluation. Otherwise, terms from the gold standard cannot be aligned, which will affect the evaluation result. Some alignment approaches or tools require a minimal frequency of the terms: this fact should be reflected in the gold standard data and used corpora.

2. Some alignment approaches align only a specific type of terms (e.g. context-based alignment only single-word terms with each other), which means that other equivalence types (e.g. single-word ↔ multi-word: (DE) Windenergie ↔ (EN) wind energy) have to be ignored in the evaluation of these specific approaches.

3. Typically, entries in the gold standard are lemmatized. However, different tokenization and lemmatization tools may process data in different ways, so that the data in the gold standard do not readily fit the corpus and/or the tool output. In such cases, manual adjustment of the gold standard is required in order to ensure that all source language terms from the gold standard can be aligned, and subsequently correctly identified in the alignment output.

4. The size of the gold standard has a big impact on the resulting evaluation scores: small lists may lead to artificially high scores which can be misleading, since each correctly aligned test term contributes massively to the final evaluation score (e.g. 1 in 20 equals 5%). Such scores are hardly comparable with scores computed on bigger data sets.

5. Evaluation metrics express different capabilities of an alignment tool: low/high number of aligned terms, number of correctly aligned terms, quality of ranking of the alignment candidates, etc.

In this poster, we show the evaluation setup we used to evaluate the alignment approaches developed in the project. We will describe problems which we faced during the evaluation and discuss a new evaluation protocol which copes with these problems.

References

---

1 www.ttc-project.eu, FP7/2007-2013, grant agreement n° 248005