Collecting Polish–German parallel corpora in the Internet

Monika Rosińska (Adam Mickiewicz University)

Bilingual corpora have recently become indispensable resources in multilingual natural language processing. Manual preparation of a bilingual corpus is a laborious task. Therefore methods for the automated creation of parallel corpora are currently a topic of concern for many researches. Firstly, an appropriate source of texts must be chosen. It has been verified by P. Resnik, M. Liberman, X. Ma and others that one of the most promising sources for obtaining appropriate texts is the Internet because of its vast size, dynamic nature, availability and multilinguality. A number of sophisticated algorithms for collecting parallel texts from the Internet have already been created. The most commonly used are STRAND[1], BITS[2] and PTMiner[3]. STRAND is the algorithm created by Philipp Resnik based on the assumption that corresponding texts have alike URL addresses and similar web page layouts (HTML tags architecture). The first version of STRAND used search engine (like AltaVista) to find candidate parallel pages. Than it verified each website language and looked for candidate pairs in the set of collected web pages (candidate pair of pages written in different languages should have similar URL, similar size). In the last stage STAND verified if texts in the candidate pair are parallel by comparing the HTML structure. In 2003 Resnik came to the conclusion that his assumptions are too general and enriched STRAND with content-based matching functions (the source code is unpublished yet). The BITS algorithm uses complex methods to define the parallelism between texts in two languages. The most efficient one uses the Like It[4] distance – it is a refined form of the edit distance applied to determine the similarity between two texts (rather than words). Firstly the texts have to be converted into bipartite graph, then the similarity is calculated as a minimum cost bipartite matching. The aim of the research has been to verify the efficiency of existing algorithms for the collection of Polish-German parallel corpora, intended as a reference source for a Machine Translation system, and possibly, to propose a new algorithm – best suitable for the task. Firstly, we developed a crawler (a program that visits consequent pages in the Internet web) and used method based on STRAND (we examined similarity of URLs and HTML tags structure) to filter candidates for parallel texts. The results were not satisfactory – only about 10% candidates were parallel pages. Therefore we decided to combine STRAND with methods inspired by the LikeIt distance applied in BITS (a source code of LikeIt has not been published) and stemming algorithms developed by us for the Polish language. To verify similarity of candidate pair we have checked permutations of numbers, and occurrence of dates. This improved results significantly – the accuracy increased to about 40%. Still, the results are clearly worse than those achieved by Resnik using pure STRAND for English and Basque or English and French. The experiment draws us to the following conclusions:

– The accuracy of existing algorithms used for language pairs like Polish and German, where neither of languages is leading in the Internet (like English), is much lower than for recently tested language pairs.

– Collection of Polish-German parallel corpora is hindered by the information noise appearing in German translations of Polish texts. The information noise has strong impact on all linguistic methods used in the algorithm. The methods must be well adjusted to overcome the problem.

– The starting web links strongly influence the efficiency of a crawler.

---

[1] "The web as parallel corpus" Philip Resnik, Noah A. Smith.

[2] "BITS: A Method for Bilingual Text Search over the Web" Xiaoyi Ma, Mark Y. Liberman.

[3] Chen, Jiang and Jian-Yun Nie. 2000. Web parallel text mining for Chinese English cross-language information retrieval. In International Conference on Chinese Language Computing, Chicago, Illinois.

[4] "The LikeIt intelligent string comparison facility" Peter N. Yianilos, Kirk G. Kanzelberger.