Hybrid algorithm for sentence alignment of Hungarian-English parallel corpora

Krisztina Tóth, Richárd Farkas, and András Kocsor

Parallel corpora, i.e. texts available in two or more languages with their segments aligned, are playing an increasingly important role in natural language processing. The processing of parallel corpora facilitates more easier for example, the building of automated machine translation systems and multilingual lexica. This paper describes an efficient hybrid technique for aligning sentences with their translation in a parallel corpus.

The Algorithm

The accuracy of a sentence-aligning algorithm depends to a great extent on the preceding sentence segmentation. Thus we proposed a sentence segmenting algorithm that works both on Hungarian and English texts and has an error rate of 0.5% or less.

In the process of alignment the sentence segmenting algorithm divides the English and Hungarian raw texts first into paragraphs, then into sentences. The output of the alignment is a TEI-compliant file containing the identifiers of the aligned Hungarian and English sentences.

In the sentence aligning algorithm we soon realised that sentences could be not extended beyond paragraph boundaries. The majority of the sentence-aligning algorithms are based on a kind of length-measure [6].

We chose the number of characters as the length-measure, and also took into consideration the fact that the number of characters within sentences have a correlation, thus in general a long sentence corresponds to a long sentence and a short sentence to a short one [3]. Obviously, it happens only in the simplest cases that a single source sentence corresponds to a single target sentence. It might happen, due to the translator's freedom of interpretation, that one source sentence corresponds to several possible target sentences and that several source sentences correspond to a single target sentence. For this reason we present a novel algorithm that is capable of recognising 1-N, N-1 and N-M correspondences as well.

We combined statistical, length-based alignment with an anchor-searching method that forms the basis of partial text-alignment. We gain such lexical information via anchor-recognition, the latter of which is present both in the target and source sentences. The anchor-searching processes published for Hungarian use the normalised form of numbers and words starting with capital letters as anchors. We added acronyms and abbreviations as anchors and employ named entities (NE) instead of words that start with a capital letter. We extracted the named entities via a quasi language-independent NE recogniser built by FARKAS et al [2]. The reason why we omitted the use of words starting with capital letter as an anchor was to avoid filtering anchors moreover, we gain valuable information by retaining inflected, lower-case named entities characteristic of agglutinative languages.

During the implementation of our hybrid algorithm we made use of both the statistical and anchor methods to examine similarities between the source and target sentences. Our costs are calculated on the basis of the length-proportion of corresponding sentences and costs assigned to the applied synchronising categories [6], just like the Gale and Church method used as a reference algorithm. We set the tolerance limit of length-proportion to 0.85 since Hungarian texts are generally 15% longer than their English equivalents. This data came from the sentence-length statistics of an English-Hungarian translation memory. We also assigned costs to similarities calculated on the basis of anchor-information, then the costs were totalled. Our cost thus derives from the costs of length-proportions measures and the costs of anchor-distances. In the sentence-alignment process the matches (mappings) with a minimal cost are chosen.

Experimental results

The implemented sentence segmenting algorithm was tested on a randomly selected part of the Szeged Treebank [1] with a result of 99.7-99.85%. This result surpasses earlier results published for the Hungarian language [4, 5].

The testing of the hybrid algorithm described above was carried out on the English-Hungarian parallel corpus built by the Research Group on Artificial Intelligence of the Hungarian Academy of Science at the University of Szeged. This corpus is comprised of 8,000 sentence pairs. The corpus contains mostly bilingual newspaper articles and general descriptions. The corpus tries to represent the Hungarian and English everyday languages. Due to the use of anchors the hybrid algorithm identified the synchronising units correctly, thus we were able to achieve more accurate results than with the length-based alignment of the Gale and Church method used as the reference algorithm. In the cases of alignment without anchors, this was determined solely by the cost of length-based matching (mapping).

References

[1] D. Csendes, J. Csirik, and T. Gyimóthy. The Szeged Corpus: A POS tagged and Syntactically Annotated Hungarian Natural Language Corpus, In *Proc. of TSD 2004*, Brno, LNAI vol. 3206, pp. 41-49, 2004.

[2] R. Farkas, Gy. Szarvas, and A. Kocsor. Named Entity Recognition for Hungarian Using Various Machine Learning Algorithm, In *Acta Cybernetica* submitted paper, 2006.

[3] W.A. Gale and K.W. Church. A Program for Aligning Sentences in Bilingual Corpora, In *29th Annual Meeting of the Association for National Linguistics*, 1991.

[4] Hunglish cd-rom

[5] A. Mihaczi, L. Németh, and M. Rácz. Magyar szövegek természetes nyelvi előfeldolgozása, In: *Magyar Számítógépes Nyelvészeti Konferencia*, pp. 40., 2003.

[6] G. Pohl. Szinkronizációs módszerek, hibrid bekezdés- és mondatszinkronizációs megoldás, In: *Magyar Számítógépes Nyelvészeti konferencia*, pp. 254-259, 2003.