

Acquis Communautaire sentence alignment using Support Vector Machines

Alexandru Ceaușu, Dan Ștefănescu, Dan Tufiș
Research Institute for Artificial Intelligence of the Romanian Academy
13, Calea 13 Septembrie, 050711, Bucharest
{alceausu, dstef, tufis}@racai.ro

Introduction

Sentence alignment is a prerequisite for any parallel corpora processing and has been proven that very good results can be obtained with practically no prior knowledge about the concerned languages. However, as the sentence alignment errors may be detrimental to further processing, ensuring higher sentence alignment accuracy is a continuous concern for many NLP practitioners. Whenever one can use additional knowledge sources about the corpora to be sentence aligned, there are good chances to improve the results. The paper describes such an aligner, which was used for the sentence alignment of the recently released 20-languages Acquis Communautaire parallel corpus (<http://wt.jrc.it/lt/acquis/>).

Related work

One of the best-known algorithms for aligning parallel corpora (Gale and Church, 1991) is based on the lengths of sentences being reciprocal translations and a very popular implementation is the Vanilla aligner (<http://nl.ijs.si/telri/Vanilla/>) due to P. Danielsson and D. Ridings. Chen (1993) developed a method based on optimizing word translation probabilities that has better results than the sentence-length based approach, but it demands much more time to complete and requires more computing resources. Melamed (1996) also developed a method based on word translation equivalence and geometrical mapping. Moore (2002) presents a hybrid approach that has three stages. In the first stage, the algorithm uses length-based methods for sentence alignment. In the second stage, a translation equivalency table is estimated from the corpus aligned in the first stage. The final step uses a combination of length-based methods and word correspondence to find 1-1 sentence alignments.

The Sentence Aligner

We describe a hybrid sentence aligner that has four stages: (I) length and geometric based sentence alignment; (II) estimation of the translation model; (III) length, geometric and word-translation based sentence alignment; (IV) recovery of the non 1-1 sentence alignments. The aligner makes use of some specific features of the Acquis Communautaire parallel corpus. The most important is the fact that the corpus has the same numbers of articles in each language. In addition, the documents observe, irrespective of the language, a precise structuring (encoded as a unique DTD).

The aligner, using a Support Vector Machine classifier, does not have language specific information and its parameters are trained using just a small portion of human checked alignment data (200 examples of correctly aligned pairs and another 200 examples of wrongly aligned pairs). In this phase of development we used an out-of-the-box solution for SVM training and classification - LIBSVM (Fan et al., 2005) with the default parameters (C-SVC classification and radial basis kernel function).

Although not strictly needed, a preliminary alignment using the hard delimiters (in our case “articles”) ensures a much faster processing.

(I) Length and geometric based sentence alignment

The first stage of our approach consists in training the SVM model on a Gold Standard that comprises 400 samples out of which half are positive examples. There are two main

features for a sentence pair: the difference in length and the difference in their relative position in the document (known as the distance from the main diagonal). We also added features like difference in character length, number of dates, number of formulas and numbers, number of punctuation marks etc.

In the next step, based on difference in relative position several pairs of alignment candidates are generated. The SVM classifier uses the trained model to discriminate the correct candidates from the erroneous ones.

(II) Estimation of the translation equivalents table

The alignments from the first stage are used to estimate the table of translation equivalents. For parameter estimation, we employed a slightly modified version of IBM model 1 (Brown et al. 1993). In the estimation, besides translation equivalency, we also use features dependent of the context of the alignment (Tufiş et al. 2005). The link locality feature accounts for the degree of the cohesion of links surrounding the candidate link. The link locality is computed for a window of words, the span of which is dependent on the aligned sentences length. Another feature we use in parameter estimation is the crossed links score that computes (for a window size also depending on the sentences lengths) the links that were crossed by the candidate link.

The parameter estimation phase of our aligner employs different weights and thresholds for features in each iteration. The weights and thresholds are manually set in order to favour the alignment of anchor words in first iterations.

(III) Length, geometric and word-correspondence based sentence alignment

In this final phase, we employ the same procedure as in the initial length and geometric based sentence alignment, but this time improved with the translation equivalents feature. This leads to a new SVM model used to discriminate between the correct and incorrect candidates for alignment.

(IV) As the above phases produce only 1-1 sentence alignments, this phase generates for the unaligned sentences (using the same alignment parameters), null alignments or multiple sentences (which should be adjacent) alignments. In our corpus we found only 0-1, 1-0, 2-1, 1-2 and 2-2 links. The preliminary analysis shows a very high accuracy of the alignments produced by our aligner. In the final paper, we will provide a detailed comparison with the alignments produced by Vanilla and Moore's aligner and an evaluation of the differences found between the three aligners.

References

- [Moore 2002] Moore, Robert C., Fast and Accurate Sentence Alignment of Bilingual Corpora. In *Machine Translation: From Research to Real Users* (Proceedings, 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Springer-Verlag, Heidelberg, Germany, 2002, 135-244
- [Gale and Church 1991] Gale, W.A., Church, K.W.: A program for Aligning Sentences in Bilingual Corpora. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California, 1991, 177-184
- [Melamed 1996] Melamed, I.D.: A Geometric Approach to Mapping Bitext Correspondence. IRCS Technical Report 96-22, University of Pennsylvania, 1996
- [Chen, 1993] Chen, S.F.: 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, 1993 9-16
- [Brown et al. 1993] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, vol. 19, 263-311, June 1993
- [Tufiş et al. 2005] D. Tufiş, Al. Ceaşu, R. Ion, D. Ştefănescu, An integrated platform for high-accuracy word alignment, JRC Enlargement and Integration Workshop: Exploiting parallel corpora in up to 20 languages, Arona, Italy, 2005
- [Fan et al, 2005] R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using the second order information for training SVM. Technical report, Department of Computer Science, National Taiwan University, 2005