[CICLING 2010: 11th International Conference on Intelligent Text Processing and Computational Linguistics, March 21-27, 2010, Iaşi, Romania]

Abstract:

Lieve Macken and Walter Daelemans. A chunk-driven bootstrapping approach to extracting translation patterns

We present a linguistically-motivated sub-sentential alignment system that extends the intersected IBM Model 4 word alignments. The alignment system is chunk-driven and requires only shallow linguistic processing tools for the source and the target languages, i.e. part-of-speech taggers and chunkers. We conceive the sub-sentential aligner as a cascaded model consisting of two phases. In the first phase, anchor chunks are linked based on the intersected word alignments and syntactic similarity. In the second phase, we use a bootstrapping approach to extract more complex translation patterns. The results show an overall AER reduction and competitive F-Measures in comparison to the commonly used symmetrized IBM Model 4 predictions (intersection, union and grow-diag-final) on six different text types for English-Dutch. More in particular, in comparison with the intersected word alignments, the proposed method improves recall, without sacrificing precision. Moreover, the system is able to align discontiguous chunks, which frequently occur in Dutch.