Identifying Word Translations Using Parallel and Comparable Corpora

Automatically generating bilingual dictionaries from parallel, i.e. manually translated texts, is an established technique that works well in practice. However, parallel texts are a scarce resource. Therefore, it is desirable to also be able to generate dictionaries from pairs of comparable monolingual corpora. For most languages, such corpora are much easier to acquire, and often in considerably larger quantities. An implemented algorithm for the identification of word translations from monolingual corpora and its results will be presented. Based on the assumption that the co-occurrence patterns between different languages are related, it expands a small base lexicon. For improved performance, an outlook will be given on some work in progress concerning a novel interlingua approach: If comparable corpora of more than two languages are available, the translations from language A to language B can not only be determined directly, but also indirectly via an intermediating language C. Given a multitude of possible language combinations, the interlingua approach promises an almost unlimited potential of cross-validation.