Abs-209

Lucia Specia and Wilker Aziz (Research Group in Computational Linguistics, University of Wolverhampton, UK)

Using a parallel corpus to learn semantic correspondences between two languages

Parallel corpora have been serving as the basis to extract different types of information for language processing applications. Common uses include extracting probabilistic word (Brown et al., 1990) or phrase dictionaries (Koehn et al., 2003) for statistical machine translation and multilingual information retrieval, and learning syntactic transfer rules for machine translation (Hoang and Koehn, 2010). In this paper we propose a method to exploit parallel corpora in order to learn legitimate shallow semantic correspondences between the two languages. An English-Spanish parallel corpus is first processed to produce shallow semantic representations. We concentrate on semantic role labels, since these can be produced automatically with satisfactory accuracy. Semantic labels, along with base-phrase information, are used to generate shallow semantic 'trees'. These are then used to learn a model of the expected correspondences in term role labels between English sentences and translations into Spanish.

References

Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. Computational Linguistics 16(2), 79-85 (1990)

Hoang, H. and Koehn, P.: Improved translation with source syntax labels. Workshop on Statistical Machine Translation and MetricsMATR, p. 409–417 (2010)

Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. pp. 48-54 (2003)