Feature-rich unsupervised word alignment models

Wilker Aziz (w.aziz@uva.nl)

Word alignment models are the very foundation of most of the research in statistical machine translation. These models infer correspondences between words in translation data as an attempt to explain how translation comes about. Translation data, otherwise known as *parallel corpus*, consist of a collection of sentences, each of which is paired with its translation in a foreign language. On the one hand, parallel data are produced by international news companies and multilingual governments as part of their routine, crucially, these data are typically freely available on the web. Word alignments, on the other hand, are an attempt to explicate translation in terms of small reusable partial translation decisions. As a consequence, we cannot easily find readily available examples of word aligned data from which to learn statistical models of word alignment. Instead, we typically face word alignment as a problem of unsupervised learning, or learning with incomplete data.

Brown et al. (1993) introduced a series of generative conditional models for word alignment (popularised as *the IBM models*), most of which (more than two decades later) still exhibit state-of-the-art performance. IBM models are estimated by the principle of maximum likelihood via the expectation maximisation (EM) technique. In EM, we use our model to "complete the data", that is, conjecture a distribution of potential explanations of the data, and then update our model as to maximise the probability it assigns to the observed data. In our case, the observed data are the parallel sentences, while the missing (hidden/latent) parts are the word alignments.

IBM models are built upon very strict independence assumptions mostly there to render a tractable EM procedure, particularly with respect to the completion step (E-step). Thus, relaxing these assumptions as to incorporate richer linguistic features is hard. Log-linear models allow the incorporation of arbitrary features into a model, however, EM training of log-linear models is intractable. An important advance is *contrastive estimation* (Smith and Eisner, 2005), by which a log-linear model can be trained on unlabelled data by approximating the E-step. Alternatively, instead of learning a set of categorical distributions directly by EM, Berg-Kirkpatrick et al. (2010) extend the intuition behind EM to feature-rich models by using multi-class logistic regression to model each categorial distribution independently.

In this project, a strong student will investigate feature-rich unsupervised models for word alignment with a particular focus on languages morphologically richer than English (e.g. German, Czech, Finnish, Turkish, Arabic, etc.), for which current word alignment models perform poorly.

References

Berg-Kirkpatrick, T., Bouchard-Côté, A., DeNero, J., and Klein, D. (2010). Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590, Los Angeles, California. Association for Computational Linguistics.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

Smith, N. A. and Eisner, J. (2005). Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 354–362, Ann Arbor, Michigan. Association for Computational Linguistics.