Machine Translation and Morphologically-rich Languages: Research Workshop of the Israel Science Foundation, University of Haifa, Israel, 23-27 January, 2011

Michael Elhadad:
Topic Models for Morphologically Rich Languages and their Usage to Explore Multilingual Corpora

Topic Models are statistical models which capture co-occurrence patterns of words across documents. Observation of topic models learned over large textual corpora indicates that they capture thematically coherent word distributions. In recent years, topic models have emerged as a useful method to infer practical semantic features for a variety of applications. In this talk, we present a variant of the Latent Dirichlet Allocation (LDA) method for morphologically rich languages and discuss its possible usage for multilingual corpus analysis.

LDA is a generative model: it assumes that documents are generated from a distribution over topics and in turn, topics are distributions over the vocabulary. Statistical inference discovers the topics that best explain a corpus. Traditionally, LDA operates over a token-based representation of documents - where tokens are sometimes filtered to remove frequent words or to keep only nouns. We show that such a token-based document representation is inadequate when applied to a Hebrew corpus, as it fails to capture coherent word patterns, due to the large number of morphological variants. We then present a lemma-based model that infers an additional layer between documents and tokens. This model successfully captures rich topic models on a variety of corpora in Hebrew, in the legal and medical domains.

Finally, we survey recent techniques used to infer multilingual topic models over multilingual corpora - both in the case of aligned and unaligned texts. We also discuss a setting where documents in Hebrew are annotated using an English ontology in the medical domain, and how the Hebrew topic models can be aligned with terms of the English ontology, thus deriving a useful resource for further machine translation processing.

This is joint work with Meni Adler, Yoav Goldberg and Rafi Cohen.