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

Kemal Oflazer:
Syntax-to-Morphology Mapping in Factored SMT

Morphology in many morphologically complex languages encodes information that is expressed with syntax involving a series of function words in a language like English. In such cases, factored phrase-based SMT is not directly applicable as there is no relation between the morphological structures of words on both sides. In this talk we present a scheme to employ factored phrase-based SMT when the morphological structures of the source and target languages are widely disparate, and experiment with it in SMT between English and Turkish. Our approach relies on syntactic analysis on the morphologically poor side (English) and then encodes a wide variety of local and non-local syntactic structures as \emph{complex structural tags} which appear as additional factors in the training data. On the morphologically-complex side (Turkish), we only perform morphological analysis and disambiguation but treat the complete complex morphological tag as a factor, instead of separating morphemes (as is traditionally done with morphologically complex languages.) Such a representation has three important side effects: (i) the length of the English sentences reduce substantially as many function words in the source are now encoded in complex tags; (ii) the abstraction of syntax mediated by function words as complex tags, allows the standard phrase extraction schemes to conflate discontinuous and continuous variants of the phrases; and (iii) since most syntax is now abstracted leaving behind mostly content words and complex tags, word alignment can operate on just root words, without noise from function words.

We incrementally explore capturing various syntactic substructures as complex tags on the English side, and evaluate how our translations improve in BLEU scores in English-to-Turkish SMT. We also apply the approach in the reverse direction but with less success, as the syntactic transformations on the English side have to be undone after SMT and this introduces some additional complications. Currently, the syntax-to-morphology transformations are manually developed and we are investigating on extracting these transformation automatically in an incremental fashion and will briefly present our ideas on this.

This is joint work with Reyyan Yeniterzi of CMU- LTI.