Statistical MT with Syntax and Morphology: Challenges and Some Solutions

Alon Lavie

Carnegie Mellon University

`alavie@cs.cmu.edu`

Phrase-based Statistical Machine Translation is the most dominant approach to MT in recent years. Its linguistic shallowness, however, limits its capabilities when applied to morphologically-rich languages and to language-pairs with highly divergent syntax. Integration of morphological analysis and syntactic modeling within statistical MT are currently at the forefront of MT research. This talk will overview recent work within my research group on hybrid MT frameworks that incorporate syntax and morphology into statistical translation.

The talk will focus on three main lines of work: (1) Morphological segmentation of Arabic and its impact on English-to-Arabic phrase-base SMT; (2) Learning of syntax-based synchronous context-free grammars from large volumes of parsed parallel corpora; and (3) Automatic Category Label Coarsening for Syntax-based MT.