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

Alon Lavie:
The Impact of Arabic Morphological Segmentation on Broad-Coverage English-Arabic Statistical MT

The relative morphological richness and complexity of Arabic in comparison with English imposes challenges on the design of phrase-based SMT systems between the two languages. Morphological segmentation impacts multiple components and resources along the MT training pipeline and runtime decoder. We explore the full spectrum of segmentation schemes ranging from full word forms to fully-segmented morphemes, and examine their impact, all within the context of training broad-scale phrase-based SMT systems between English and Arabic. Our results and analysis indicate a complex picture, where some morphological decomposition choices help improve performance whereas others have no impact or hurt performance. Furthermore, the resulting variation in system output can be further leveraged for improving performance via system combination.