

Issues in the generation of Arabic

The exceptional degree of ambiguity in the writing system, the rich morphology, the unique word formation process of roots and patterns, and its syntax all contribute to making computational approaches to Arabic very challenging. The non-concatenative morphology of Arabic has spurred the development of sophisticated formalisms and computational engines, as well as produced brute force approaches. Widely different computational approaches to this special morphological system have been proposed. From the perspective of the knowledge-based approach, the main challenge is the necessity of linearizing the non-linear. In other words, we need to concatenate what is non-concatenative at some level of morphological knowledge representation. In this presentation, we address the issue of where the linearization process should take place: in the lexicon component, in the morphological component or partially in both. We weigh these options, based on Arabic data and linguistic evidence. We also describe a set of crucial problems we have encountered in the generation of Arabic sentences in an English-to-Arabic knowledge-based machine translation (MT) system.

A demonstration of the morphological and the sentence generators, using examples of the major issues, will be shown.