Statistical Machine Translation into a Morphologically Complex Language

Kemal Oflazer

Faculty of Engineering and Natural Sciences
Sabancı University
Istanbul, Tuzla, 34956, Turkey
oflazer@sabanciuniv.edu

Abstract. In this paper, we present the results of our investigation into phrase-based statistical machine translation from English into Turkish – an agglutinative language with very productive inflectional and derivational word-formation processes. We investigate different representational granularities for morphological structure and find that (i) representing both Turkish and English at the morpheme-level but with some selective morpheme-grouping on the Turkish side of the training data, (ii) augmenting the training data with "sentences" comprising only the content words of the original training data to bias root word alignment, and with highly-reliable phrase-pairs from an earlier corpus-alignment (iii) re-ranking the n-best morpheme-sequence outputs of the decoder with a word-based language model, and (iv) "repairing" translated words with incorrect morphological structure and words which are out-of-vocabulary relative to the training and the language model corpus, provide an non-trivial improvement over a word-based baseline despite our very limited training data. We improve from 19.77 BLEU points for our word-based baseline model to 26.87 BLEU points for an improvement of 7.10 points or about 36% relative. We briefly discuss the applicability of BLEU to morphologically complex languages like Turkish and present a simple extension to compare tokens not in a all-or-none fashion but taking lexical-semantic and morpho-semantic similarities into account, implemented in our BLEU+ tool.

1 Introduction

Statistical machine translation from English-to-Turkish poses a number of difficulties. Typologically English and Turkish are rather distant languages: while English has very limited morphology and rather fixed SVO constituent order, Turkish is an agglutinative language with a very rich and productive derivational and inflectional morphology, and a very flexible (but SOV dominant) constituent order. One implication of complex morphology is that, in parallel texts, Turkish words usually align to multiple words on the English side. When done at the word level, this is very noisy and masks the more (statistically) meaningful alignments at the sub-lexical level. Another issue of practical significance is the lack of large scale parallel text resources, with no substantial improvement expected