Introduction

In this study, the TÜBİTAK-UEKAE statistical machine translation system based on the open-source phrase-based statistical machine translation software, Moses, with added components to address the rich morphology of the source languages is presented. Additionally, 3 submissions (primary, contrastive 1, contrastive 2) which use:
- linguistic morphological analysis and statistical disambiguation to generate morpheme-based translation models;
- unsupervised subword segmentation to generate non-linguistic subword-based translation models;
- and word-based models but makes use of lexical approximation to cope with out-of-vocabulary words, respectively.

We describe the preprocessing and postprocessing steps and our training and decoding procedures.

Coping with Turkish Morphology

Turkish is an agglutinative language where words can carry several morphemes in the form of suffixes. E.g. Morphological decomposition of the Turkish word and the morpheme-based alignment to its English translation:

- 'slowly'  
- 'quickly'  
- 'tightly'  
- 'if you will not be able to do'

Statistical machine translation involving Turkish requires special attention to Turkish morphology.

Three approaches to dealing with the morphology of Turkish are investigated:

- Development of a morphological analyzer requires bits of manual work and linguistic expertise.
- An unsupervised morphological analyzer, called Morfessor is used.
- Morfessor uses the minimum description length (MDL) principle to find an optimal subword segmentation of a given corpus in the form of a root-and-morpheme vocabulary.

The segmentation in this model are static in that all the occurrences of a word are assumed to be segmented in the same manner regardless of the context.

Unsupervised Morphological Segmentation

- Development of a morphological analyzer requires bits of manual work and linguistic expertise.
- An unsupervised morphological analyzer, called Morfessor is used.
- Morfessor uses the minimum description length (MDL) principle to find an optimal subword segmentation of a given corpus in the form of a root-and-morpheme vocabulary.
- The segmentation in this model are static in that all the occurrences of a word are assumed to be segmented in the same manner regardless of the context.

Lexical Approximation

- As an alternative to morphological segmentation, we investigated the usefulness of the lexical approximation approach we had previously used in WMT 2007.
- The corpus and the translation models remain word-based; however, a morphological analyzer may be utilized internally to compute a similarity feature between words based on their shared roots and morphemes.