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 unsupervised subword segmentation to generate morpheme-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:

  - yap +a +ma +yacak +sa +n
  - yap: to do
  - a: be
  - ma: able
  - yacak: to be
  - sa: able
  - n: to not

- Statistical machine translation involving Turkish requires special attention to Turkish morphology.
- Three approaches to dealing with the morphology of Turkish are investigated:
  1. Development of a morphological analyzer requires bits of manual work and linguistic expertise.
  2. An unsupervised morphological analyzer, called Morfessor is used.
  3. Finite-state morphological analyzer by Kemal Oflazer and statistical disambiguator of Sak et al. are used.

Unsupervised Morphological Segmentation

- 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 segmentations 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

- Similar to our 2007 and 2008 systems, we made use of phrase table augmentation.

Results

- Among the three morphological approaches for Turkish, using morphological analysis customized to the translation task performed the best (primary submission).
- The word-based lexical approximation approach performed close to unsupervised segmentation, even though it was outperformed during development experiments.
- In our experiments, training a segmentation model for the English side and using it in system training did not provide a clear improvement over leaving the English corpus as words.
- The added complexity of generating roots and morphemes at the decoder output, and the errors in English morphological segmentation could be reasons.

Table below shows the effect of N-gram model order on the performance of our primary submission.

<table>
<thead>
<tr>
<th>LM Order</th>
<th>Arabic-English</th>
<th>Turkish-English</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.512</td>
<td>0.5102</td>
</tr>
<tr>
<td>2</td>
<td>0.5340</td>
<td>0.5328</td>
</tr>
<tr>
<td>3</td>
<td>0.5565</td>
<td>0.5554</td>
</tr>
<tr>
<td>4</td>
<td>0.5771</td>
<td>0.5761</td>
</tr>
<tr>
<td>5</td>
<td>0.5977</td>
<td>0.5967</td>
</tr>
<tr>
<td>6</td>
<td>0.6183</td>
<td>0.6173</td>
</tr>
<tr>
<td>7</td>
<td>0.6381</td>
<td>0.6371</td>
</tr>
<tr>
<td>8</td>
<td>0.6577</td>
<td>0.6567</td>
</tr>
<tr>
<td>9</td>
<td>0.6775</td>
<td>0.6765</td>
</tr>
<tr>
<td>10</td>
<td>0.6972</td>
<td>0.6962</td>
</tr>
</tbody>
</table>

- We used the open-source statistical machine translation toolkit Moses for training the translation models and for decoding.
- An N-gram English language model was trained using the SRI language modeling toolkit.
- All the system training and decoding was performed on lowercased and punctuation-tokenized data.
- Although we used 3-gram target language models in our systems, Table below shows the effect of N-gram model order on the performance of our primary submission.

- Similar to our 2007 and 2008 systems, we made use of phrase table augmentation.