Addressing Problems across Linguistic Levels in SMT: Combining Approaches to Model Morphology, Syntax and Lexical Choice

Marion Weller-Di Marco1,2  Alexander Fraser2  Sabine Schulte im Walde1
1IMS – University of Stuttgart  2CIS – Ludwig-Maximilians-University of Munich
{dimarco|schulte}@ims.uni-stuttgart.de  fraser@cis.lmu.de

**PROBLEMS ACROSS LINGUISTIC LEVELS**

**Structural differences in source/target language**
- Often difficult to capture with word alignment
- Long-distance reordering is typically costly during translation

**Lexical choice**
- Word sense disambiguation, selectional preferences, ...
- Translation of multi-word expressions

**Morphological complexity**
- Data sparsity due to uncovered inflected forms
- Difficulty to produce the correct target-side inflection based on available information

**SYNTACTIC LEVEL**
- English verbs are moved to the expected German position
  - moving the verb to verb-final position
  - moving the verb to verb-second position
- Resulting structure considerably different from "regular" English

**LEXICAL LEVEL**
- Sentence-level source-side features, target features restricted to phrase
- Features for discriminative model
  - pos/word/lemma (source-side window, target-side phrase)
  - dependency information (source-side)

**MORPHOLOGICAL LEVEL**
- Inflection prediction process to handle nominal morphology:
  - translation in stemmed representation
  - generation of inflected forms
- Markup with translation relevant inflectional features

**COMBINING APPROACHES**

- **Pre-processing – syntactic level**
  - Source-side reordering
- **At decoding time – lexical level**
  - Discriminative classifier to score translation rules using source-side context
- **Post-processing – morphological level**
  - Target-side inflection prediction

→ EN–DE phrase-based SMT system
→ Do individual gains add up when combining approaches?
→ Are there side-effects between the linguistic levels?

**RESULTS**
- EN–DE phrase-based Moses system
- 4.5 M parallel sentences
- 5-gram language model (45 M sentences)

**Morpho-syntactic + lexical strategies**

<table>
<thead>
<tr>
<th>system</th>
<th>basic</th>
<th>VW1 - pos/lem</th>
<th>VW2 - pos/lem/dep</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface</td>
<td>19.45</td>
<td>19.81</td>
<td>19.90</td>
</tr>
<tr>
<td>Surface V-Reordered</td>
<td>19.71</td>
<td>20.24</td>
<td>20.27</td>
</tr>
<tr>
<td>MorphSys</td>
<td>19.81</td>
<td>19.90</td>
<td>19.93</td>
</tr>
<tr>
<td>MorphSys V-Reordered</td>
<td>20.08</td>
<td>20.51</td>
<td>20.50</td>
</tr>
</tbody>
</table>

**Annotating number + tense information**

<table>
<thead>
<tr>
<th>system</th>
<th>VW2</th>
<th>+num</th>
<th>+num+tense</th>
</tr>
</thead>
<tbody>
<tr>
<td>MorphSys</td>
<td>19.93</td>
<td>20.00</td>
<td>20.02</td>
</tr>
<tr>
<td>MorphSys V-Reordered</td>
<td>20.50</td>
<td>20.57</td>
<td>20.62*</td>
</tr>
</tbody>
</table>

**Examples for Number and Tense Features**

<table>
<thead>
<tr>
<th>SRC reordered</th>
<th>i really feel that <strong>he</strong> should follow in the footsteps of the other guys.</th>
</tr>
</thead>
<tbody>
<tr>
<td>+NumTense</td>
<td>ich bin wirklich der Meinung, dass er in die Fußstapfen der anderen Jungs folgen soll.</td>
</tr>
<tr>
<td>VW2 reordered</td>
<td>es wäre daher geeignet, die illegale Einwanderung in die USA zu unterstützen.</td>
</tr>
<tr>
<td>+NumTense</td>
<td>es wäre daher ideal, illegale Einwanderung in die USA zu unterstützen.</td>
</tr>
</tbody>
</table>

**Conclusion**
- Combination of established approaches to address the three linguistic levels: *Morphology, Syntax and Lexical Choice*
- The strategies are complementary
- Reordered systems benefit more from discriminative model
- Additional features aiming at verbal inflection lead to further improvement