We present an English-German SMT system that deals with complex target-side morphology by applying a two-step translation process: (cf. [2],[3])

- translation model built on stems;
- prediction of morphological features, generation of inflected forms.

Improving case prediction

- Due to the flexible German clause ordering, case is difficult to predict.
- Case is an important indicator of the role of an NP in the sentence; the most difficult is to distinguish – syntactic functions (subject, direct/indirect object) – modifying NPs (genitive modification).

New features for case prediction

- projection of source-side syntactic information; information about target-side syntactic frames obtained from dependency-parsed corpora.

### 2. Overview of the inflection process

#### Morphological Features

- The gender of an NP is part of the stem.
- English input determines the number of an NP.
- Strong/weak inflection depends on the choice of determiner and the setting of the other features.
- There are 4 values for case: nominative (Subject), accusative (direct object), dative (indirect object) and genitive (modification, object in rare cases).

#### Feature prediction and inflection

- Individual sequence models for each morph.
- The models have access to stems, POS-tags within a window of four positions.
- Generate inflected forms using features and stems: bleu[4D]<>nose[<fem><sg><weak>] -> blue (cf. [1]).

#### 3. Motivation for case modeling

**Source-side features**

1. (do) unterstellen
2. (that) unterstützt
3. (the) Minister
4. (has) agreed
5. (the) report
6. (has) submitted
7. (the) report
8. (has) given

**Subcategorization information**

- External knowledge base comprises dependency information on verbal subcategorization frames.
- We model the association strength for verb-noun constructions: subject, benefactive, patient.
- berth (report) is more likely to be patient (direct object) than Mitarbeiter (employee).
- zustimmen (agree) has a preference for only selecting subject and indirect object theme.
- The remaining NP cannot receive case from the verb and is thus a genitive modifier of the NP-pair.

#### 4. Subcategorization features

**Verb-noun tuples with case information**

<table>
<thead>
<tr>
<th>Tuple</th>
<th>Gloss</th>
<th>Case Values</th>
<th>Nouns Acc</th>
<th>case- features</th>
</tr>
</thead>
<tbody>
<tr>
<td>haben</td>
<td>die</td>
<td>die</td>
<td>the</td>
<td>government</td>
</tr>
<tr>
<td>Anordnen</td>
<td>der</td>
<td>der</td>
<td>of the</td>
<td>United States</td>
</tr>
<tr>
<td>Regierung</td>
<td>Staat</td>
<td>Staat</td>
<td>State</td>
<td>United States</td>
</tr>
<tr>
<td>Minister</td>
<td>Minister</td>
<td>Minister</td>
<td>minister</td>
<td>Mr.</td>
</tr>
<tr>
<td>Buses</td>
<td>Mittel</td>
<td>Mittel</td>
<td>Middle</td>
<td>Financial</td>
</tr>
<tr>
<td>vom</td>
<td>aus</td>
<td>aus</td>
<td>from</td>
<td>Financial</td>
</tr>
</tbody>
</table>

**N-N pairs with frequency information**

<table>
<thead>
<tr>
<th>Tuple</th>
<th>Gloss</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>blaue</td>
<td>Entwicklung</td>
<td>492</td>
</tr>
</tbody>
</table>

**Integration of source-side and subcategorization features**

- Extraction of verb-noun tuples and candidates for N-Ngen constructions:
  - based on syntactic trees produced by a hierarchical SMT-system
  - derived from source-side dependencies via word alignment
- Look up co-occurrence probabilities/frequencies.

**5. Integration of source-side and subcategorization features**

- German dependency relations are transferred to the SMT output based on the word alignment.
- Information about the complete tuple (verb+noun and N-Ngen) is annotated as bigram, e.g. Regierung+anordnen.

### 6. Experiments and evaluation

<table>
<thead>
<tr>
<th>System</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>44</td>
<td>45</td>
<td>47</td>
<td>48</td>
</tr>
</tbody>
</table>

Table: BLEU scores for different inflections (1-4).

- Hierarchical SMT system using GKM target-side syntax trained on WMT-2009 data (En/De).
- Inflection prediction better than surface system.
- Systems 1-4: different inflections of the same SMT output; system 1 does not use new features.
- No significant difference between the enriched systems and the simple prediction system.
- No changes in stem sequence, but different inflection; BLEU can hardly capture the difference.

#### Manual evaluation

- Human annotators prefer the enriched system.

### 7. Conclusion

- We presented a two-step SMT system that translates into stems and generates inflected forms.
- We illustrated the need for external knowledge sources to model case and presented a translation system using source-side syntactic features and a subcategorization database.
- A manual evaluation showed that the proposed features have a positive impact.
- First integration of explicit subcat-information from large monolingual corpora into SMT.

### 8. References