Predicting Prepositions for SMT

Marion Weller1,2, Alexander Fraser2, Sabine Schulte im Walde1
1University of Stuttgart, 2Ludwig Maximilian University of Munich

1. Introduction & Motivation

Translating prepositions is difficult in SMT
- Convey the source-side meaning
- Meet target-side requirements

How are prepositions generated in translation?
- Functional prepositions: determined by target-side requirements
to believe in sth. → an etw glauben
to learn from sth. → von etw lernen [person] etc.
to etw. lernen [abstract]
- Content-bearing prepositions: largely determined by source-side preposition
to sit ander/on the table → unter/au dem Tisch sitzen
- “In-between”: source- and target-side play a role
go in the cinema/to the beach → ins Kino/an den Strand gehen

Modeling prepositions on the target-side
- Abstract representation during translation in a morphology-aware EN-DE SMT system
- Generation of prepositions as post-processing step

2. Modeling Prepositions

Subcategorization: difficult to capture in SMT
- Grammatical case of noun phrases corresponds to the syntactic function (subject, direct/indirect object)

Objective: model all subcategorized elements (PP/subject/object) of a verb
- All arguments available in an abstract form
- Are then assigned their respective functional overt preposition → PP
- empty preposition → NP
- Arguments are then inflected accordingly

Realization of prepositions is independent of structural mismatches of source/target side
to pay attention to sth. → auf etw. achten
to etw. beachten
- both variants are possible, but require a different realization of the preposition depending on the verb

Overview of the translation process
(1) Building the morphology-aware SMT system
- lemmatization for representation
- Target-side prepositions are replaced with place-holders;
  “empty” place-holders are inserted at the beginning of NPs
- “empty” prepositions added to source-side NPs
(2) Generating surface forms
- prediction and realization of place-holder prepositions as overt preposition (PP) or “empty” preposition (NP)
- prediction of inflection-relevant morphological features
  - generation of inflected forms

3. Overview: Translation and Prediction Steps

<table>
<thead>
<tr>
<th>input</th>
<th>lemmatized SMT output</th>
<th>prep.</th>
<th>morph. features</th>
<th>inflected</th>
<th>gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>PREP</td>
<td>R-PREP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>what</td>
<td>welch&lt;NPAT&gt;</td>
<td>Acc</td>
<td>Acc.Fem,Sg,Wk</td>
<td>welche</td>
<td>which</td>
</tr>
<tr>
<td>role</td>
<td>Rolle&lt;NN&gt;+&lt;Fem&gt;+&lt;Sg&gt;</td>
<td>Acc</td>
<td>Acc.Fem,Sg,Wk</td>
<td>Rolle</td>
<td>role</td>
</tr>
<tr>
<td>@</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>d ie&lt;ART&gt;</td>
<td>Nom</td>
<td>Nom.Masc,Sg,St</td>
<td>der</td>
<td>the</td>
</tr>
<tr>
<td>giant</td>
<td>riesig&lt;ADJ&gt;</td>
<td>Nom</td>
<td>Nom.Masc,Sg,St</td>
<td>design</td>
<td>giant</td>
</tr>
<tr>
<td>planet</td>
<td>Planet&lt;NN&gt;+&lt;Masc&gt;+&lt;S g&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>has</td>
<td>gepli ekt&lt;VPVP&gt;</td>
<td>-</td>
<td>gespielt</td>
<td>played</td>
<td></td>
</tr>
<tr>
<td>played</td>
<td>hat&lt;VAFIN&gt;</td>
<td>be-Dat</td>
<td></td>
<td>hat</td>
<td>has</td>
</tr>
<tr>
<td>in</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>die&lt;ART&gt;</td>
<td>Dat</td>
<td>Dat.Fem,Sg,Wk</td>
<td>der</td>
<td>the</td>
</tr>
<tr>
<td>development</td>
<td>Ent wicklung&lt;NN&gt;+&lt;Fem&gt;+&lt;Sg&gt;</td>
<td></td>
<td>Entwicklung</td>
<td>development</td>
<td></td>
</tr>
<tr>
<td>of</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td>die&lt;ART&gt;</td>
<td>Gen</td>
<td>Gen.Neut,Sg,St</td>
<td>des</td>
<td>of-the</td>
</tr>
<tr>
<td>solar system</td>
<td>Son nensystem&lt;NN&gt;+&lt;Neut&gt;+&lt;Sg&gt;</td>
<td></td>
<td>Sonensystem</td>
<td>solar system</td>
<td></td>
</tr>
</tbody>
</table>

German Cases: Nominative – subject; Accusative – direct object; Dative – indirect object; Genitive – nominal modifier

4. Features for Predicting Prepositions

- Target-side context: adjacent lemmas/POS tags
- Source-side features
  - signed word on source-side: overt or empty preposition
governed noun and its syntactic function to its governor
governing verb or noun of source-side preposition
- Projected source-side features
  - governing target verb, governed target noun
- Distributional subcategorization preferences
  - information in form of e.g. verb-preposition-case tuples
  - learn, whether a given combination predominantly occurs as subject, direct/indirect object, PP or noun/noun modification
- Prediction models: CRFs trained with Wapiti
  - Prediction accuracy: 73.5% (prep+case); 85.7% (prep)

5. Prediction Features in the Training Data

<table>
<thead>
<tr>
<th>lemma</th>
<th>gloss</th>
<th>source-side</th>
<th>prp</th>
<th>func.noun</th>
<th>g.verb</th>
<th>projected-source-side</th>
<th>g.verb</th>
<th>target-side</th>
<th>subcat</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>aber</td>
<td>but</td>
<td>-</td>
<td>-</td>
<td>subj, we</td>
<td>endure</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PREP</td>
<td>prep</td>
<td>0</td>
<td>subj</td>
<td>we</td>
<td>endure</td>
<td>leiden</td>
<td>N-Nom:5</td>
<td>@-Acc@ unter-Dat:4</td>
<td>@-Nom</td>
<td>-</td>
</tr>
<tr>
<td>wir</td>
<td>we</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>leiden</td>
<td>leiden</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>-</td>
<td>-</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>auch</td>
<td>too</td>
<td>-</td>
<td>-</td>
<td>obj</td>
<td>effect</td>
<td>Treibhaus-effekt</td>
<td>N-Nom:5</td>
<td>@-Acc@ unter-Dat:4</td>
<td>@-Nom</td>
<td>-</td>
</tr>
<tr>
<td>PREP</td>
<td>prep</td>
<td>0</td>
<td>subj</td>
<td>we</td>
<td>endure</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>die</td>
<td>Treib haus</td>
<td>green house</td>
<td>-</td>
<td>effect</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Source sentence with inserted empty prepositions: ... , @ we too are having to endure @ the green house effects

6. Abstract Representation of Prepositions

- “Basic” place-holder → decreased translation quality
  - S1 Plain place-holders
  - Enriched abstract representation
    - S2 Grammatical case
      - S3 Governing target verb, governed target noun
    - S4 Functional vs. content-conveying
      - S5 Assuming that functional prepositions convey less in terms of meaning
      - Replace functional prepositions with place-holders
      - keep “regular” prepositions for content-conveying prepositions

7. Experiments

- Standard phrase-based Moses system
- 4.3M parallel EN-DE sentences, 10.3M lines LM-data
- Test/tuning sets: 3000 sentences news data

<table>
<thead>
<tr>
<th>system</th>
<th>Prepositions BLEU LMT</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>16.84</td>
<td>-</td>
</tr>
<tr>
<td>baseline++</td>
<td>17.38</td>
<td>-</td>
</tr>
</tbody>
</table>

- No improvement over baseline; best result obtained with annotation of case (S2)

Automatic evaluation of generated prepositions
- Subset where relevant parts (governed noun, governing verb) match with the reference
- No real improvement over baseline

Example
- “It is also important to learn how to improve your own performance.”
  - Translation: Das ist auch wichtig zu lernen, wie man die eigene Leistung zu verbessern.
  - Result: 16.11

8. Conclusion & Future Work

- Generation of prepositions based on an abstract representation using source and target features
  - handle structural differences between source/target-side
- No improvement over morphology-aware baseline
  - annotation of grammatical case → best system

How to improve the current method?
- Abstract representation
  - grammatical case: light semantic annotation
  - obtain a more meaningful representation by more
  - semantically motivated annotation to represent the class of a preposition (temporal, local, directional, ...)
- Integration of the generation step
  - integrate into decoding process

9. Selected Related Work

Agree et al. (2006) use an MSA-lemmatizer to Transfer Prepositions and Determiners from Arabic to English
Weller et al. (2016) generate prepositions appropriately for respective context, but translation without place-holder representation