Delimiting Morphosyntactic Search Space with Source-Side Reordering Models

Joachim Daiber, Khalil Sima’an

Institute for Logic, Language and Computation
University of Amsterdam
Motivation

- Current MT models work well if languages are structurally similar
- Difficulties with morphologically rich languages:
  - freer word order
  - more productive morphological processes
  - agreement over long distances
Motivation

"Germans like to buy holiday homes in Florida"

- Deutsche kaufen sich meistens in Florida eine Ferienwohnung
- Deutsche kaufen sich in Florida meistens eine Ferienwohnung
- In Florida kaufen sich meistens Deutsche eine Ferienwohnung
- In Florida kaufen sich Deutsche meistens eine Ferienwohnung
- Meistens kaufen sich Deutsche in Florida eine Ferienwohnung
- ...

From: Frankfurter Allgemeine Zeitung (August 31, 2015)
Motivation

“Germans like to buy holiday homes in Florida”

- Deutsche kaufen sich meistens in Florida eine Ferienwohnung
- Deutsche kaufen sich in Florida meistens eine Ferienwohnung
- In Florida kaufen sich meistens Deutsche eine Ferienwohnung
- In Florida kaufen sich Deutsche meistens eine Ferienwohnung
- Meistens kaufen sich Deutsche in Florida eine Ferienwohnung
- ...

From: Frankfurter Allgemeine Zeitung (August 31, 2015)
Preordering source trees

Source dependency trees are good fit for preordering:

- Lerner and Petrov (2013) present two classifier-based dep. tree preordering models
- Jehl et al. (2014) and de Gispert et al. (2015) preorder dep. trees via branch-and-bound search
Preordering source trees

- Lerner and Petrov (2013) preorder trees starting at the root
- Order all children (model 1) or left and right children (model 2)
Preordering source trees

- Lerner and Petrov (2013) preorder trees starting at the root
- Order all children (model 1) or left and right children (model 2)
Preordering source trees

- Lerner and Petrov (2013) preorder trees starting at the root
- Order all children (model 1) or left and right children (model 2)
Preordering source trees

- Lerner and Petrov (2013) preorder trees starting at the root
- Order all children (model 1) or left and right children (model 2)
Generating the space of potential word order choices

- Both Lerner and Petrov (2013) and Jehl et al. (2014) make only *single-best* predictions
- We want:
  - *ALL REASONABLE* predictions instead of *SINGLE BEST*
  - More flexible model
Producing multiple predictions

Multiple predictions:

- Bad: Mistakes in order decisions propagate
- Extract $n$-best decisions from the model to pass to subsequent model
Producing multiple predictions

Model over possible orders of source words:

\[ P(s' \mid s, \tau) = \prod_{h \in \tau} P_T(\pi_h \mid s, h, \tau) \]
Producing multiple predictions

Model over possible orders of source words:

\[ P(s' \mid s, \tau) = \prod_{h \in \tau} P_T(\pi_h \mid s, h, \tau) \]

- Preordered \( s \)
Producing multiple predictions

Model over possible orders of source words:

\[ P(s' \mid s, \tau) = \prod_{h \in \tau} P_T(\pi_h \mid s, h, \tau) \]

- Preordered \( s \)
- Source dep. tree
Producing multiple predictions

Model over possible orders of source words:

\[
P(s' | s, \tau) = \prod_{h \in \tau} P_T(\pi_h | s, h, \tau)
\]

- Preordered \( s \)
- Source dep. tree
- Heads of all families
Producing multiple predictions

Model over possible orders of source words:

\[ P(s' \mid s, \tau) = \prod_{h \in \tau} P_T(\pi_h \mid s, h, \tau) \]

- Preordered \( s \)
- Source dep. tree
- Heads of all families
- Local permutation
Producing multiple predictions

Model over possible orders of source words:

\[ P(s' | s, \tau) = \prod_{h \in \tau} P_T(\pi_h | s, h, \tau) \]

\[ P_T(\pi | s, h, \tau) = P(\psi | s, h, \tau) P_L(\pi_L | s, h, \tau) P_R(\pi_R | s, h, \tau) \]
Producing multiple predictions

Model over possible orders of source words:

\[ P(s' \mid s, \tau) = \prod_{h \in \tau} P_T(\pi_h \mid s, h, \tau) \]

\[ P_T(\pi \mid s, h, \tau) = P(\psi \mid s, h, \tau) \quad P_L(\pi_L \mid s, h, \tau) \quad P_R(\pi_R \mid s, h, \tau) \]

- Pivot decision
Producing multiple predictions

Model over possible orders of source words:

\[ P(s' \mid s, \tau) = \prod_{h \in \tau} P_T(\pi_h \mid s, h, \tau) \]

\[ P_T(\pi \mid s, h, \tau) = P(\psi \mid s, h, \tau) \quad P_L(\pi_L \mid s, h, \tau) \quad P_R(\pi_R \mid s, h, \tau) \]

- Pivot decision
- Left order decision
Producing multiple predictions

Model over possible orders of source words:

\[ P(s' \mid s, \tau) = \prod_{h \in \tau} P_T(\pi_h \mid s, h, \tau) \]

\[ P_T(\pi \mid s, h, \tau) = P(\psi \mid s, h, \tau) P_L(\pi_L \mid s, h, \tau) P_R(\pi_R \mid s, h, \tau) \]

- Pivot decision
- Left order decision
- Right order decision
Producing multiple predictions

Model over possible orders of source words:

\[ P(s' \mid s, \tau) = \prod_{h \in \tau} P_T(\pi_h \mid s, h, \tau) \]

\[ P_T(\pi \mid s, h, \tau) = P(\psi \mid s, h, \tau) \quad P_L(\pi_L \mid s, h, \tau) \quad P_R(\pi_R \mid s, h, \tau) \]
Producing multiple predictions

Model over possible orders of source words:

\[
P(s' \mid s, \tau) = \prod_{h \in \tau} P_T(\pi_h \mid s, h, \tau)
\]

\[
P_T(\pi \mid s, h, \tau) = P(\psi \mid s, h, \tau) \quad P_L(\pi_L \mid s, h, \tau) \quad P_R(\pi_R \mid s, h, \tau)
\]
Preordering algorithm

- Produce $k_P$ best pivot decisions for all the children in the family
- For every of the $k_P$ pivot decisions:
  - Produce $k_L$ best left order decisions
  - Produce $k_R$ best right order decisions
Preordering with arbitrary non-local features

Making the model more flexible:

- Bad: Order decisions are local to tree families
- Khalilov and Sima’an (2012) show even weak LM helps with shortcomings
Preordering with arbitrary non-local features

Decoding:
- Non-local features ruin our day...
- Cube pruning to the rescue (Chiang, 2007)!
Preordering with arbitrary non-local features

Preordering model:

- Standard log-linear model (Och and Ney, 2002):

  \[
  \hat{s'} = \arg \max_{s'} \sum_i \lambda_i \log \phi_i(s')
  \]

- Where to get the weights?
  - PRO: tuning as ranking (Hopkins and May, 2011)
  - Scoring functions:
    1. Kendall’s \( \tau \) coefficient
    2. Simulate word level MT system, score by BLEU
Preordering with arbitrary non-local features

Local features:
- Lexicalized preordering model $P(s' | s, \tau)$ from before
- Unlexicalized preordering model $P_W(\pi | h, cs)$ as less sparse backoff

Non-local features:
- ngram language models over $s'$
  - words
  - part-of-speech tags
  - word classes
Applicability of this model

- General model is applicable to any $n$-best preordering model over source trees

- Example:
  - Preordering model: Pairwise neural network-based model (de Gispert et al., 2015)
  - Parsing algorithm: $k$-best ITG-based CKY parsing (similar to Tromble and Eisner (2009)).
Intrinsic: Do non-local features help?

- Intrinsic evaluation of preordering quality
- Language pair English-to-German

<table>
<thead>
<tr>
<th>Model</th>
<th>Kendall’s tau</th>
<th>BLEU ($s' \rightarrow s'$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First-best −LM</td>
<td>92.16</td>
<td>68.1</td>
</tr>
<tr>
<td>First-best +LM (cube)</td>
<td>92.27</td>
<td>68.7</td>
</tr>
</tbody>
</table>
Translation: Quality of potential word order choices

- Translation experiments with the space of word order choices
- Experiments with top 10 preordering outputs of this model

<table>
<thead>
<tr>
<th></th>
<th>Distortion</th>
<th>BLEU</th>
<th>MTR</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>7</td>
<td>15.20</td>
<td>35.43</td>
<td>66.62</td>
</tr>
<tr>
<td>Best out of $k$ ($k = 10$)</td>
<td></td>
<td>17.26*</td>
<td>37.97*</td>
<td>62.64</td>
</tr>
</tbody>
</table>
Discussion

Preordering with non-local features

- Integration of LM helps improve preordering quality
  - Slight Kendall $\tau$ improvement
  - BLEU preorder score shows benefits mostly in small local windows

Quality of the space of potential word order choices

- Experiments show significant potential improvement contained in the space
- With arbitrary $n$ or lattice, space is small enough to be handled by subsequent models
Conclusion

- Source preordering has big limitations but has proven very successful
- Our interest: Source-side adaptation models more suitable for morphologically rich languages

- First steps towards this goal:
  - Introduced preordering model that can delimit space instead of first-best predictions
  - More flexible model with arbitrary non-local features and cube pruning
Thank You!

Any questions?


