Max-Margin Synchronous Grammar Induction for Machine Translation

Xinyan Xiao and Deyi Xiong
Soochow University
Synchronous Grammar Induction

Obama hold a talk with Netanyahu

Obama $X_1 \rightarrow$ Obama $X_1$

with Netanyahu

$X_1$ hold a talk $X_1$

......
Word-based Heuristics (Chiang, 2007)

<table>
<thead>
<tr>
<th>Obama</th>
<th>hold</th>
<th>a</th>
<th>talk</th>
<th>with</th>
<th>Netanyahu</th>
</tr>
</thead>
<tbody>
<tr>
<td>奥巴马</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>与</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>内塔尼亚胡</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>举行</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>会谈</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Word-based Heuristics  (Chiang, 2007)

<table>
<thead>
<tr>
<th>Obama</th>
<th>hold</th>
<th>a</th>
<th>talk</th>
<th>with</th>
<th>Netanyahu</th>
</tr>
</thead>
<tbody>
<tr>
<td>奥巴马</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>与</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>内塔尼亚胡</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>举行</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>会谈</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$X_1$举行会谈 → hold a talk $X_1$
Generative Model (Levenberg et al., 2012)

- Max-likelihood
- Hard to integrate features

References:
- Marcu and Wong, 2002
- Cherry and Lin, 2007
- Zhang et al., 2008
- DeNero et al., 2008
- Blunsom et al., 2009
- Cohn and Blunsom, 2009
- Neubig et al., 2011
- Levenberg et al., 2012
Discriminative Model (Xiao et al., 2012)

- Max-likelihood
- Only local Feature
  - Source Parse Structure

Obama hold a talk with Neta.
This Work

- Max-margin
- 1 - BLEU
- Non-local feature
  - Target Parse Structure
Discriminative Model

- **Scoring Function**
  \[ f(s, t, d) = \theta^T \varphi(s, t, d) \]

- **Feature Function**
  \[ \varphi(s, t, d) = \sum_{r \in d} \varphi(r, s) + \sum_{r \in d} \varphi(r, s, t) \]
  Local
  Non-local
Obama hold a talk with Neta.
Max-margin Estimation

- Margin is large than smoothed sentence BLEU-4

\[ f(s^{(i)}, t^{(i)}) - f(s^{(i)}, t) \geq 1 - \text{BLEU-4}(t^{(i)}, t) \]
Optimization

For each sent.

- Biparse Reference
- Collect Rule
- Cost-Augmented Viterbi
- Biparse Viterbi
- Update Weights
Optimization

For each sent.

Biparse Reference

Collect Rule

Cost-Augmented Viterbi

Biparse Viterbi

Update Weights

Obama hold a talk with Neta.
For each sent.

- Biparse Reference
- Collect Rule
- Cost-Augmented Viterbi
- Biparse Viterbi
- Update Weights

Optimization

Obama hold a talk with Neta.

X_1 行 会谈 → hold a talk X_1
Optimization

For each sent.

- Biparse Reference
- Collect Rule
- Cost-Augmented Viterbi
- Biparse Viterbi
- Update Weights

\[ f(s^{(i)}, t) \rightarrow \text{BLUE-4}(t^{(i)}, t) \]
Optimization

For each sent.

1. Biparse Reference
2. Collect Rule
3. Cost-Augmented Viterbi
4. Biparse Viterbi
5. Update Weights

Obama with Neta. hold a talk
Optimization

For each sent.

Biparse Reference

Collect Rule

Cost-Augmented Viterbi

Biparse Viterbi

Update Weights

(1) Sub-gradient: reference - viterbi
(2) Projection: rescale weights

Shalev shwartz et al. (2007)
Experiment

- Bilingual data: 110M sentence pairs from LDC
- 5-gram LM: 432M words from LDC
- Dev: MT02
- Test: MT03, 04, 05
  - Report average BLEU scores on the three test data
- Comparing systems
  - Moses-chart
  - Baseline: In-house implementation of Hiero
Comparison against Traditional Pipeline

- +1.3 BLEU over Baseline

- Sparse Feature: rule, phrase boundary, orientation boundary
Comparison against Traditional Pipeline

- 23% rules over baseline
Comparison against Max-likelihood

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Max-likelihood</th>
<th>Max-margin +non-local Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>31.5</td>
<td>32.5</td>
<td>33.5</td>
</tr>
</tbody>
</table>
Conclusion

- Max-margin grammar induction framework
  - First work: Learn and optimize synchronous grammar towards BLEU
  - Incorporate non-local features from target tee
  - Outperform both traditional pipeline and max-likelihood method

- Future direction
  - Apply our framework on linguistically syntax-based system
  - Incorporate contextual model
Thank you!
Backup
Cube Pruning based Biparsing

- Create k-best hyperedges for each source span from the Bottom-up
  - Enumerate all inferable source parses, and create cubes.
- Cube pruning
Factorize Hyperedge

- Hyperedge can be factorized into smaller structures
- Construct them by combining sub-structures
Find a Hyperedge

1. Enumerate source side

2. Create cubes

3. Construct hyperedges

one of few

0.2 \times 2.5

0.2 \times 3.5
Parse a source span

- Combine all potential cubes of source span
- Create k-best hyperedges for all cubes