A Topic Similarity Model for Hierarchical Phrase-based Translation

Xinyan Xiao, Deyi Xiong, Min Zhang, Qun Liu, Shouxun Lin

2012-7-10
Topic is Important for Translation

Bank

Mouse
Topic is Important for Translation

Bank

Finance

Geography

Mouse

Biology

Computer

老鼠

河岸
Previous Work

- **Topic-specific lexicon translation**
  - Focus on word-to-word translation
  - Estimate $P(e|f, \text{topic})$

- **State-of-the-art systems translate by sequence of rules**
  - A rule is rarely factorized into individual words

---

Zhao and Xing 2007, Tam et al., 2007
This Work: Rule Topic Distribution

给与 X\(_1\) ⇔ give X\(_1\)

作战 能力 ⇔ operational capacity

给予 X\(_1\) ⇔ grants X\(_1\)

X\(_1\) 举行 会谈 X\(_2\) ⇔ held talks X\(_1\) X\(_2\)
This Work: Similarity and Sensitivity

- **Topic Similarity Model**
  - Describe the relatedness of rules to topics of given documents

- **Topic Sensitivity Model**
  - Distinguish topic-insensitive rules and topic-sensitive rules
Topic Similarity Model

Source Document
Topic Similarity Model

Source Document
Topic Similarity Model

\[ \sum_{k=1}^{K} \left( \sqrt{\hat{p}(z = k|d)} - \sqrt{\hat{p}(z = k|r)} \right)^2 \]
Topic Sensitivity Model

- **Topic-insensitive Rule**: Applied in many topics

- **Topic-sensitive Rule**: Applied in few topics

- Describe by **Entropy** as a metric
Topic Sensitivity Model

- Topic-insensitive rules are always penalized
- But common, sometime more preferable
- Sensitivity as a complement
Estimation
Estimation

\[(r_1, 0.1, \ldots) \quad (r_2, 0.5, \ldots) \quad \ldots \ldots\]
Estimation

$$(r_1, 0.1, \ldots) \quad (r_2, 0.5, \ldots) \quad \ldots$$
One-to-many Topic Projection

Topic Assignment Alignment

Topic-to-Topic Projection Matrix

\[
\begin{array}{cccc}
f_{10} & f_{15} & f_{10} & f_{26} \\
e_3 & & & \\
e_3 & & & \\
e_8 & & & \\
e_8 & & & \\
\end{array}
\]

\[
\begin{bmatrix}
0.1 & 0.4 & 0.1 & \cdots \\
0.3 & 0.2 & 0.1 & \cdots \\
0.4 & 0.1 & 0.3 & \cdots \\
\ldots & \ldots & \ldots & \ldots \\
\end{bmatrix}
\]
One-to-many Topic Projection

Topic Assignment Alignment

Topic-to-Topic Projection Matrix

Target Distribution

Projected Target Distribution
Decoding

\[ \lambda_1 \text{Similarity}(\text{source}), + \lambda_2 \text{Similarity}(\text{target}) + \lambda_3 \text{Sensitivity}(\text{source}) + \lambda_4 \text{Sensitivity}(\text{target}) + \sum \lambda_k h(f, e) \]
Experiment Setup

- In-house implementation of HPB model
- Topic Tool: GibbsLDA++
- Bilingual corpus: FBIS 239K sentence pairs
  - With document boundary
  - For both LDA training and rule extraction
- Report Average BLEU on test sets NIST06, NIST08
Effect of Topic Similarity Model

- Baseline
- Topic Lex
- SimSrc
- SimTgt
- SimSrc+SimTgt
- Sim+Sem

Values:
- Baseline: 25.8
- Topic Lex: 26.0
- SimSrc: 26.2
- SimTgt: 26.4
- SimSrc+SimTgt: 26.6
- Sim+Sem: 26.8
Effect of Sensitivity Model

- Baseline
- Topic Lex
- SimSrc
- SimTgt
- SimSrc+SimTgt
- Sim+Sem

Comparison of models with varying sensitivity.
# One-to-many Topic Projection

<table>
<thead>
<tr>
<th>e-topic</th>
<th>f-topic 1</th>
<th>f-topic 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>enterprises</td>
<td>农业 (agricultural)</td>
<td>企业 (enterprise)</td>
</tr>
<tr>
<td>rural</td>
<td>农村 (rural)</td>
<td>市场 (market)</td>
</tr>
<tr>
<td>state</td>
<td>农民 (peasant)</td>
<td>国有 (state)</td>
</tr>
<tr>
<td>agricultural</td>
<td>改革 (reform)</td>
<td>公司 (company)</td>
</tr>
<tr>
<td>market</td>
<td>财政 (finance)</td>
<td>金融 (finance)</td>
</tr>
<tr>
<td>reform</td>
<td>社会 (social)</td>
<td>银行 (bank)</td>
</tr>
<tr>
<td>$P(z_f</td>
<td>z_e)$</td>
<td>0.38</td>
</tr>
</tbody>
</table>

![Graph showing baseline, one-to-one, and one-to-many projections with corresponding P(z_f|z_e) values.](image-url)
Conclusion

- Exploiting topic model at the rule level is better than at the word level
- Both similarity model and sensitivity model help translation
- Future Directions
  - Apply on data without document boundary
  - Extend to other translation models
谢谢
Thank you
Topic Similarity Model

- Context-based translation

He et al., 2008; Shen et al., 2009

Source Document

给予 $X_1 \Rightarrow$ give $X_1$

给予 $X_1 \Rightarrow$ grants $X_1$
Topic Sensitivity Model

Given $X_1 \Rightarrow give X_1$

Topic-insensitive Rule
14.9%

Topic-sensitive Rule
85.1%

- Average Entropy of Document Topic Distribution
Effect on Various Types of Rules

- Baseline
- Phrase
- Mono
- Reorder
- All

Effect measured on a scale from 25.8 to 27.