A Neural Reordering Model for Phrase-based Translation

Peng Li
Tsinghua University
pengli09@gmail.com

joint work with Yang Liu, Maosong Sun, Tatsuya Izuha, Dakun Zhang
Phrase-based Translation

布什 与 沙龙 举行 了 会谈

Bush held a talk with Sharon

segmentation  reordering  translation

(Koehn et al., 2003; Och and Ney, 2004)
Q: Can you figure out a sentence using these words?
Chinese President Xi Jinping and his Us counterpart Barack Obama open two days of talks in California on a number of high-stakes issues

Q: Can you figure out a sentence using these words?
Reordering is Hard

• An NP-complete problem (Knight, 1999; Zaslavskiy et al., 2009)

• Reordering modeling has attracted intensive attention, e.g.

  • Distance-based model (Koehn et al., 2003)

  • Word-based lexicalized model (Koehn et al., 2007)

  • Phrase-based lexicalized model (Tillman, 2004)

  • Hierarchical phrase-based lexicalized model (Galley and Manning, 2008)
Distance-based Model

布什 与 沙龙 举行 了 会谈
Bush held a talk with Sharon

(Koehn et al., 2003)
Lexicalized Models

(Koehn et al., 2007; Tillman, 2004; Galley and Manning, 2008)
Lexicalized Models

布什 与 沙龙 举行 了 会谈

Bush held a talk with Sharon

(Koehn et al., 2007; Tillman, 2004; Galley and Manning, 2008)
Lexicalized Models

布什 与 沙龙 举行 了 会谈

Bush held a talk with Sharon

(Koehn et al., 2007; Tillman, 2004; Galley and Manning, 2008)
# Challenge #1: Sparsity

<table>
<thead>
<tr>
<th>Source Phrase</th>
<th>Target Phrase</th>
<th>M</th>
<th>S</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>布什</td>
<td>Bush</td>
<td>0.7</td>
<td>0.2</td>
<td>0.1</td>
</tr>
<tr>
<td>举行 了 会谈</td>
<td>held a talk</td>
<td>0.1</td>
<td>0.1</td>
<td>0.8</td>
</tr>
<tr>
<td>与 沙龙</td>
<td>with Sharon</td>
<td>0.7</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>举行 了</td>
<td>held a</td>
<td>0.6</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>会谈</td>
<td>talk</td>
<td>0.4</td>
<td>0.3</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Challenge #1: Sparsity

- Probability distributions are estimated by MLE.
Challenge #2: Ambiguity

(a) ti gao xin yong ka ying yun de tou ming du
...... 提高 信用卡 营运 的 透明度， ......

...... enhanced transparency of credit card business.

(b) yi ji ying yun mo shi zhuo wei zhi zao
...... 以及 营运 模式 转为 制造 ......

...... the changing mode of business towards a more ......

(c) jin yi bu gai shan jian zhu ye de ying yun
...... 进一步 改善 建 造 业 的 营运。

...... further improve business in the construction industry.
Challenge #3: Context

Insensitivity

How to resolve the three challenges?

Bush held a talk with Sharon
Including More Contexts

Sparsity

Ambiguity

Context Insensitivity
Sparsity

• Including more contexts leads to severer sparsity

Reordering as Classification
Neural Reordering Model

- A neural classifier for predicting reordering orientations
- Conditioned on both the current and previous phrase pairs
  - Improves context sensitivity
  - Reduces reordering ambiguity
- A single classifier for all phrase pairs
  - Uses vector space representations
  - Alleviates the data sparsity problem
Recursive Autoencoder (RAE)

\[
[x'_1; x'_2] = f^{(2)}(W^{(2)}y_1 + b^{(2)})
\]

\[
y_1 = f^{(1)}(W^{(1)}[x'_2; x'_1] + b^{(1)})
\]

\[
\|x_1 - x'_1\|^2
\]

(Pollack; 1990; Socher et al, 2011)
Recursive Autoencoder (RAE)

\[ [y'_1; x'_3] = f^{(2)}(W^{(2)} y_2 + b^{(2)}) \]

\[ \|y_1 y_2 y'_1\|^2 (W^{(1)} [y_1; x_3] + b^{(1)}) \]

\[ \|x_3 - x'_3\|^2 \]

(Pollack; 1990; Socher et. al, 2011)
Neural Classifier

orientations

RAE

softmax

current phrase pair

previous phrase pair

held a talk

with Sharon
Training

Reordering error on predicting orientations

Reconstruction error on recovering training examples
Reconstruction Error

- Reconstruction error

\[ E_{rec}([c_1; c_2]; \theta) = \frac{1}{2} \| [c_1; c_2] - [c'_1; c'_2] \|^2 \]

- Source side average reconstruction error

\[ E_{rec,s}(S; \theta) = \frac{1}{N_s} \sum_i \sum_{p \in T_R^6(t_i, s)} E_{rec}([p.c_1, p.c_2]; \theta) \]

- Total reconstruction error

\[ E_{rec}(S; \theta) = E_{rec,s}(S; \theta) + E_{rec,t}(S; \theta) \]
Reordering Error

• Average cross-entropy error

\[ E_{reo}(S; \theta) = \frac{1}{|S|} \sum_i \left( -\sum_o d_{t_i}(o) \cdot \log(P_\theta(o|t_i)) \right) \]

• Joint training objective

\[ J = \alpha E_{rec}(S; \theta) + (1 - \alpha) E_{reo}(S; \theta) + R(\theta) \]

\[ R(\theta) = \frac{\lambda_L}{2} \| \theta_L - \theta_{L_0} \|^2 + \frac{\lambda_{rec}}{2} \| \theta_{rec} \|^2 + \frac{\lambda_{reo}}{2} \| \theta_{reo} \|^2 \]
Optimization

- Hyper-parameters optimization
  - $\alpha, \lambda_L, \lambda_{rec}, \lambda_{reo}$
  - Optimized by random search (Bergstra and Bengio, 2012)

- Training objective optimization: L-BFGS
  - Using backpropagation through structures to compute the gradients (Goller and Kuchler, 1996)
Experiments

• Chinese–English translation

• Training: 1.2M sentence pairs

• LM: 4-gram, 397.6M words

• Dev. set: NIST 06

• Test set: NIST 02–05, 08

• Case-insensitive BLEU

• Baselines

• Distance-based model

• Lexicalized model

\[
\begin{array}{c}
\text{word-based} \\
\text{phrase-based} \\
\text{hier. phrase-based}
\end{array}
\times
\begin{array}{c}
\text{M/S/D} \\
\text{left/right}
\end{array}
\]
M/S/D Orientations

• Care about relative position and adjacency
Left/Right Orientations

• Only care about relative position

Bush held a talk with Sharon right

Bush held a talk with Sharon left
Translation

<table>
<thead>
<tr>
<th>BLEU</th>
<th>MT06 (dev)</th>
<th>MT02</th>
<th>MT03</th>
<th>MT04</th>
<th>MT05</th>
<th>MT08</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>30.75</td>
<td>31.25</td>
<td>31.25</td>
<td>31.25</td>
<td>31.25</td>
<td>31.25</td>
</tr>
<tr>
<td>neural M/S/D</td>
<td>33.00</td>
<td>33.00</td>
<td>33.00</td>
<td>33.00</td>
<td>33.00</td>
<td>33.00</td>
</tr>
<tr>
<td>neural left/right</td>
<td>33.00</td>
<td>33.00</td>
<td>33.00</td>
<td>33.00</td>
<td>33.00</td>
<td>33.00</td>
</tr>
</tbody>
</table>
Non-Separability

- The unaligned Chinese word “de” makes a big difference in determining M/S/D orientations

Kinmen has 60000 resident population

金门有六万的常住人口
Non-Separability

- The unaligned Chinese word “de” makes a big difference in determining M/S/D orientations

金门 有 六 万 的 常住 人口
Kinmen has 60000 resident population
Non-Separability

六万的常住人口
60000 resident population

六万常住人口
60000 resident population
Non-Separability

- Left/right orientations are not so sensitive to unaligned words

Kinmen has 60,000 resident population
Non-Separability

• Left/right orientations are not so sensitive to unaligned words

金门 有 六 万 的 常住 人口

Kinmen has 60000 resident population
Non-Separability

六万的常住人口

六万的常住人口

60000 resident population

60000 resident population

left right
Non-Separability
Distortion Limit

![Distortion Limit Graph]

- Blue line: neural
- Green line: lexicalized
Word Vectors

![Bar plot showing BLEU scores for task-oriented and word2vec vectors across different datasets.](Image)
Vector Space Representations
Conclusion

• We propose a neural reordering model for phrase-based translation

• It improves the context sensitivity, reduces ambiguity and alleviates the data sparsity problem

• Future work

  • Train MT system and neural classifier jointly
  • Develop more efficient models to leverage larger contexts
  • Extend our work to syntax-based and n-gram based models
Thanks!