Recursive Autoencoders for ITG-based Translation

Peng Li
Tsinghua University
pengli09@gmail.com

(Joint work with Yang Liu and Maosong Sun)
Overview

- **Phrase reordering model** is a critical problem in machine translation (MT), and is NP-complete.

\[
\text{我 有 一个 从 没有 见 过 的 女性 朋友。}
\]

\[
\text{I have a female friend never seen before.}
\]

(Knight, 1999)
**Distortion Models**

- **Distortion models**: penalize relative displacement of source phrases

I have a female friend never seen before.

(Koehn et al., 2003; Och and Ney, 2004)
Distortion Models

- **Distortion models**: penalize relative displacement of source phrases

\[ d = 0 \]

(Koehn et al., 2003; Och and Ney, 2004)
Distortion Models

- **Distortion models**: penalize relative displacement of source phrases

I have a female friend never seen before.

(Koehn et al., 2003; Och and Ney, 2004)
Distortion Models

- **Distortion models**: penalize relative displacement of source phrases

\[ d=0 \quad d=5 \]

(Koehn et al., 2003; Och and Ney, 2004)
Distortion Models

- **Distortion models**: penalize relative displacement of source phrases

我 有 一个 从 没有 见 过 的 女性 朋友。

I have a female friend never seen before.

(Koehn et al., 2003; Och and Ney, 2004)
Distortion Models

- **Distortion models**: penalize relative displacement of source phrases

```
我有一个从没见过的女性朋友。

I have a female friend never seen before.
```

(Koehn et al., 2003; Och and Ney, 2004)
Lexicalized Reordering Models

- **Lexicalized reordering models**: penalize reordering conditioned on both the source and target phrases

  - Bush
  - hold a talk
  - with Sharon

  - 布什
  - 与 沙龙
  - 举行 了 会谈

(Koehn et al., 2007)
Lexicalized Reordering Models

- **Lexicalized reordering models**: penalize reordering conditioned on both the source and target phrases

  Bush hold a talk with Sharon

(Koehn et al., 2007)
Lexicalized Reordering Models

- **Lexicalized reordering models**: penalize reordering conditioned on both the source and target phrases

  
<table>
<thead>
<tr>
<th>布什</th>
<th>与 沙龙</th>
<th>举行 了 会谈</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bush</td>
<td>hold a talk</td>
<td>with Sharon</td>
</tr>
</tbody>
</table>

(Koehn et al., 2007)
Lexicalized Reordering Models

- Lexicalized reordering models: penalize reordering conditioned on both the source and target phrases

Bush hold a talk with Sharon

M D S

(Koehn et al., 2007)
Lexicalized Reordering Models

- **Lexicalized reordering models**: penalize reordering conditioned on both the source and target phrases

I have a female friend never seen before.

(Koehn et al., 2007)
Block Merging

- Reordering as block merging

I have a female friend never seen before.

(Wu, 1997; Xiong et al., 2006)
Block Merging

• Reordering as block merging

I have a female friend never seen before.

(Wu, 1997; Xiong et al., 2006)
Block Merging

- Reordering as block merging

I have a female friend never seen before.

(Wu, 1997; Xiong et al., 2006)
Block Merging

• Reordering as block merging

我有一个从没见过的女性朋友

I have a female friend never seen before

invert

(Wu, 1997; Xiong et al., 2006)
• Reordering as block merging

I have a female friend never seen before.

(Wu, 1997; Xiong et al., 2006)
Block Merging

- Reordering as block merging

I have a female friend never seen before.

straight

(Wu, 1997; Xiong et al., 2006)
Block Merging

- Reordering as block merging

我 有一个 从 没有 见过 的 女性 朋友

I have a female friend never seen before

(Wu, 1997; Xiong et al., 2006)
Block Merging

- Reordering as block merging

I have a female friend never seen before. (Wu, 1997; Xiong et al., 2006)
Block Merging

• Reordering as block merging

我有一个从没有见过的女性朋友。

I have a female friend never seen before.

(Wu, 1997; Xiong et al., 2006)
I have a female friend never seen before.

(Wu, 1997; Xiong et al., 2006)
I have a female friend never seen before.

(Wu, 1997; Xiong et al., 2006)
I have a female friend never seen before.

(Wu, 1997; Xiong et al., 2006)
I have a female friend never seen before.

(Wu, 1997; Xiong et al., 2006)
I have a female friend never seen before.

(Wu, 1997; Xiong et al., 2006)
Block Merging

- Can you find a counter example?
Block Merging

- Can you find a counter example?

... 进一步 就 中东 危机 举行 会谈

... hold further talk on the Mideast crisis

(Huang et al., 2009)
Block Merging

“inside-outside”

(Wu, 1997)
ITG

• Inversion transduction grammar (ITG)

\[ X \rightarrow [X^1, X^2] \quad : \text{straight rule} \]
\[ X \rightarrow \langle X^1, X^2 \rangle \quad : \text{inverted rule} \]
\[ X \rightarrow f/e \quad : \text{lexical rules} \]
ITG-based Reordering Model

• **Type I:** Incorporating ITG into left-to-right decoding to constrain the reordering space (e.g., Zens et al., 2004; Feng et al., 2010)

• **Type II:** Translation as ITG parsing, e.g.

• **Max-Ent ITG reordering model:** using maximum entropy (MaxEnt) model to predict which rule to use (Xiong et al., 2006)
MaxEnt ITG Reordering Model

Potentially alleviates the data sparseness problem

How to extract features from training examples?

• Which words are representative for predicting reordering?

• Xiong et al. (2006) only use boundary words
This Work

• We propose an ITG reordering classifier based on recursive autoencoders (RAE)

• Our model considers the whole phrases

• RAEs can produce vector space representations for arbitrary strings

• Our system achieves 1.07 BLEU points improvement on NIST 2008 dataset
Neural ITG Reordering Model

“never seen before” v.s. “seen before never”
Neural ITG Reordering Model

Real-valued vector

RAE
Neural ITG Reordering Model

straight inverted

Softmax layer
Translation
Translation
Translation
Translation
Translation
Translation
Translation
Translation
Translation
Translation
Autoencoders

• Each word is represented as a vector, e.g.
  • “female” ➤ $[0.1 \ 0.8 \ 0.4]^{\top}$
  • “friend” ➤ $[0.7 \ 0.1 \ 0.5]^{\top}$

• What is the vector representation of “female friend”?
Autoencoders

- **Encoding**
  \[ p = f^{(1)}(W^{(1)}[c_1; c_2] + b^{(1)}) \]

- **Decoding**
  \[ [c'_1; c'_2] = f^{(2)}(W^{(2)}p + b^{(2)}) \]

- **What about multi-word strings?**
Recursive Autoencoders

\[ y_3 = f^{(1)}(W^{(1)}[y_2; x_4] + b) \]

\[ y_2 = f^{(1)}(W^{(1)}[y_1; x_3] + b) \]

\[ y_1 = f^{(1)}(W^{(1)}[x_1; x_2] + b) \]

(Socher et. al, 2011)
Training

Reordering error: how well the classifier predicts the merging order?

Reconstruction error: how well the learned vector space representations represent the corresponding strings?
Reconstruction Error

• Reconstruction error

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

• Source side average reconstruction error

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

• Total reconstruction error

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

• Average cross-entropy error

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

• Joint training objective

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

\[
R(\theta) = \frac{\lambda_L}{2} \|\theta_L - \theta_{L0}\|^2 + \frac{\lambda_{rec}}{2} \|\theta_{rec}\|^2 + \frac{\lambda_{reo}}{2} \|\theta_{reo}\|^2
\]
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 gradients (Goller and Kuchler, 1996)
Experiments

- Training corpus: 1.23M sentence pairs
- Language model: 4-gram language model trained on the Xinhua portion of the GIGAWORD corpus
- Dev. set: NIST 2006 MT dataset
- Test set: NIST 2008 MT dataset
- Metric: case-insensitive BLEU-4 score
## BLEU-4

<table>
<thead>
<tr>
<th>System</th>
<th>NIST06 (dev)</th>
<th>NIST08 (tst)</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxent</td>
<td>30.40</td>
<td>23.75</td>
</tr>
<tr>
<td>neural</td>
<td>31.61*</td>
<td>24.82*</td>
</tr>
</tbody>
</table>

*: significantly better (p < 0.01)
## BLEU-4

<table>
<thead>
<tr>
<th>Sentence Length</th>
<th>&gt;</th>
<th>=</th>
<th>&lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1, 10]</td>
<td>43</td>
<td>121</td>
<td>57</td>
</tr>
<tr>
<td>[21, 30]</td>
<td>170</td>
<td>11</td>
<td>152</td>
</tr>
<tr>
<td>[31, 40]</td>
<td>105</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td>[41, 50]</td>
<td>69</td>
<td>1</td>
<td>53</td>
</tr>
<tr>
<td>[51, 119]</td>
<td>40</td>
<td>0</td>
<td>30</td>
</tr>
</tbody>
</table>
Classification Accuracy

Accuracy (%)

Length

neural

maxent
Conclusion

• We have presented an ITG reordering classifier based on RAEs

• Feature work
  • Combine linguistically-motivated labels with recursive neural networks
  • Investigate more efficient decoding algorithms
  • Apply our method to other phrase-based and even syntax-based systems
Reference


Reference


Reference


Thanks!
Backup Slides
## Training Data Size

<table>
<thead>
<tr>
<th># of examples</th>
<th>NIST06 (dev)</th>
<th>NIST08 (tst)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100,000</td>
<td>30.88</td>
<td>23.78</td>
</tr>
<tr>
<td>200,000</td>
<td>30.75</td>
<td>23.89</td>
</tr>
<tr>
<td>300,000</td>
<td>30.80</td>
<td>24.35</td>
</tr>
<tr>
<td>400,000</td>
<td>31.01</td>
<td>24.45</td>
</tr>
<tr>
<td>6,004,441</td>
<td>31.61</td>
<td>24.82</td>
</tr>
</tbody>
</table>
## Cluster Examples

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>works for verify on tunnels from transparency in opinion at</td>
<td>these people who the reasons why the story of how the system which the trend towards</td>
<td>of the three on the fundamental over the entire through its own with the best</td>
</tr>
</tbody>
</table>