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 了 沙龙 举行 会谈

Bush held a talk with Sharon

segmentation  reordering  translation

(Koehn et al., 2003; Och and Ney, 2004)
Reordering is Hard

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

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

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 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

| Source Phrase       | Target Phrase      | M  | S  | D  |
|---------------------|--------------------|----|----|----|
| 布什                | Bush               | 0.7| 0.2| 0.1|
| 举行 了 会谈         | held a talk        | 0.1| 0.1| 0.8|
| 与 沙龙              | with Sharon        | 0.7| 0.1| 0.2|
| 举行 了              | held a             | 0.6| 0.1| 0.3|
| 会谈                | talk               | 0.4| 0.3| 0.3|
Challenge #1: Sparsity

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

(a) 提高 信用卡 营运 的 透明度， ......  
...... enhanced transparency of credit card business .

(b) 以及 营运 模式 转为 制造 ......  
...... the changing mode of business towards a more ......

(c) 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

held a talk with Sharon
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)

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

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

\[ y_1 = f^{(1)}(W^{(1)}x'_2 \| x_1, x'_2 \|^2 + b^{(1)}) \]

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

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

\[ ||y_1y_2 \neq f^{(1)}(W^{(1)}[y_1; x_3] + b^{(1)}) ||^2 \]

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

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

current phrase pair

previous phrase pair
Training

Reordering error on predicting orientations

Reconstruction error on recovering training examples
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

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

• 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

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

- Joint training objective

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

\[ R(\theta) = \frac{\lambda_{L}}{2} \|\theta_{L} - \theta_{L_0}\|^2 + \frac{\lambda_{\text{rec}}}{2} \|\theta_{\text{rec}}\|^2 + \frac{\lambda_{\text{reo}}}{2} \|\theta_{\text{reo}}\|^2 \]
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
    \[
    \begin{cases}
    \text{word-based} \\
    \text{phrase-based} \\
    \text{hier. phrase-based}
    \end{cases}
    \times \begin{cases}
    \text{M/S/D} \\
    \text{left/right}
    \end{cases}
    \]
M/S/D Orientations

• Care about relative position and adjacency
Left/Right Orientations

- Only care about relative position
Translation

![Bar chart showing BLEU scores for different MT models.]

- **MT06 (dev)**
  - Baseline: 30.75
  - Neural M/S/D: 33
  - Neural left/right: 24

- **MT02**
  - Baseline: 30.75
  - Neural M/S/D: 33
  - Neural left/right: 24

- **MT03**
  - Baseline: 30.75
  - Neural M/S/D: 33
  - Neural left/right: 24

- **MT04**
  - Baseline: 30.75
  - Neural M/S/D: 33
  - Neural left/right: 24

- **MT05**
  - Baseline: 30.75
  - Neural M/S/D: 33
  - Neural left/right: 24

- **MT08**
  - Baseline: 30.75
  - Neural M/S/D: 33
  - Neural left/right: 24
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

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

Kinmen has 60,000 resident population.
Non-Separability

六万的常住人口

60000 resident population

六万的常住人口

60000 resident population

M S D
Non-Separability

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

Kinmen has 60000 resident population

金门有六万的常住人口
Non-Separability

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

```
jinmen you liu wan de changzhu renkou
金门 有 六 万 的 常住

Kinmen has 60000 resident population
```
Non-Separability

The diagram illustrates the concept of non-separability with a visual representation of 'left' and 'right' categories.
Non-Separability
Distortion Limit

![Graph showing BLEU scores against Distortion Limit for neural and lexicalized models.](image)
Word Vectors

- Task-oriented vectors
- word2vec vectors

| BLEU  | MT06  | MT02  | MT03  | MT04  | MT05  | MT08  |
|-------|-------|-------|-------|-------|-------|-------|
| 30.5  | 34    | 30.5  | 30.5  | 30.5  | 30.5  | 23.5  |
|       | MT06  | MT02  | MT03  | MT04  | MT05  | MT08  |

MT06 (dev)
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!