Translation Similarity Model Based on Bilingual Compositional Semantics

Wang Chaochao, Xiong Deyi, Zhang Min

Soochow University

December 8, 2014
Outline

1. Background
2. Motivation
3. Translation Similarity Model
4. Experiments
5. Conclusion and Future Work
Background

1. Distributional Hypothesis
   Occurring in similar contexts, having same meaning (Harris, 1968)

2. Distributional Semantics

3. Challenging and Important.

4. Helpful to phrase-based SMT

Example

<table>
<thead>
<tr>
<th>Woman</th>
<th>100</th>
<th>200</th>
<th>0</th>
<th>500</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
<td>money</td>
<td>cosmetic</td>
<td>beard</td>
<td>shopping</td>
<td>drinking</td>
</tr>
<tr>
<td>Man</td>
<td>150</td>
<td>10</td>
<td>200</td>
<td>50</td>
<td>150</td>
</tr>
</tbody>
</table>
Compositional Semantics

1. Principle
   - meanings
   - rules

2. Methods
   - $\lambda$-Calculus
   - Vector Addition
   - other deep learning methods
Why does SMT need Compositional Semantics

**Example**

- **Chinese Sentence:** 小张说故事很有趣.
  

  **Different Composition, Different Meanings.**
Why does SMT need Compositional Semantics

Example

- **Chinese Sentence:** 小张说故事很有趣.
- **Composition 1:** 小张 说 故事 很 有趣.
- **Translation 1:** Zhang said the story is very interesting.

*Different Composition, Different Meanings.*
Why does SMT need Compositional Semantics

Example

- **Chinese Sentence:** 小张说故事很有趣.
- **Composition 1:** 小张 说 故事 很 有趣.
- **Translation 1:** Zhang said the story is very interesting.
- **Composition 2:** 小张 说故事 很 有趣.
- **Translation 2:** Zhang telling story is very interesting.

*Different Composition, Different Meanings.*
**Motivation**

### Previous Approaches

1. **Word-level Semantics**
   - Mikolov et al. (2013): *Transformation Matrix*
   - Zou et al. (2013): *Bilingual Training*

2. **Phrase-level Semantics**
   - Gao et al. (2014): *Jointly Training*

### Problems

1. Still focus on **linear projection**
Motivation

Previous Approaches

1. Word-level Semantics
   - Mikolov et al. (2013): Transformation Matrix
   - Zou et al. (2013): Bilingual Training

2. Phrase-level Semantics
   - Gao et al. (2014): Jointly Training

Problems

1. Still focus on linear projection
2. Defined at word level
Motivation

Previous Approaches

1. **Word-level Semantics**
   - Mikolov et al. (2013): *Transformation Matrix*
   - Zou et al. (2013): *Bilingual Training*

2. **Phrase-level Semantics**
   - Gao et al. (2014): *Jointly Training*

Problems

1. Still focus on **linear projection**
2. Defined at **word level**
3. **Jointly learning** limits application in SMT
Our Approach

Step 1: Monolingual Compositional Semantics
Step 2: Projection
Step 3: Calculating and Integrating
Goal: \textit{Learning Compositional Representations}

How?

1. Point-Wise Mutual Information based vector space word representations

   \[ pmi(c, t) = \log \frac{p(c, t)}{p(c)p(t)} \]

2. Weighted Vector Addition

   \[ \vec{p} = \alpha \vec{w}_1 + \beta \vec{w}_2 \]
**Goal:**

*Learning Bilingual Representations*

**Project**

![Diagram of a neural network with layers labeled Input, Hidden, and Output, and nodes connecting source phrases to target phrases via weight matrices $W_1$ and $W_2$.](image)
Translation Similarity Model

Goal:
Calculating Similarity

How?

1. Between Two Phrases

\[ Sim(p(\vec{w}_1), \vec{w}_2) = \frac{p(\vec{w}_1) \cdot \vec{w}_2}{\|p(\vec{w}_1)\| \times \|\vec{w}_2\|} \]

2. Source Sentence

\[ M_{sim} = \prod_{(w_1, w_2) \in P} sim(p(\vec{w}_1), \vec{w}_2) \]
How to Integrate into SMT?

As a feature in any SMT systems.

Log Linear Model

1. General Log Linear Model

\[
Pr(e|f) \approx P_{\lambda_1...\lambda_M}(e|f) = \frac{\exp[\sum \lambda_m h_m(e, f)]}{\sum \exp[\lambda_m h_m(e', f)]}
\]

2. New Feature \( M_{sim} \)

\[
M_{sim} = \prod_{(w_1, w_2) \in P} \text{sim}(p(\overrightarrow{w_1}, \overrightarrow{w_2})}
\]
Experiment Setup

1. Language: Chinese → English
2. Baseline system: Hierarchical Phrase-based System
3. Training data: 4.1M sentence pairs from LDC data
4. Development data: NIST 2005
5. Test data: NIST 2006 and NIST 2008
6. Open Source Toolkit: DISSECT
   (To get word representations and weights)
### Results

<table>
<thead>
<tr>
<th>Systems</th>
<th>Development Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NIST 05</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>33.53</td>
<td>30.66</td>
</tr>
<tr>
<td>Non-projection</td>
<td>32.76</td>
<td>30.38</td>
</tr>
<tr>
<td>Linear Projection</td>
<td>34.62</td>
<td>30.77</td>
</tr>
<tr>
<td>Non-linear Projection</td>
<td>34.53</td>
<td><strong>31.22</strong></td>
</tr>
</tbody>
</table>

### Analysis

1. Compositional semantics can improve translation quality
# Results

<table>
<thead>
<tr>
<th>Systems</th>
<th>Development Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NIST 05</td>
<td>NIST 06</td>
</tr>
<tr>
<td>Baseline</td>
<td>33.53</td>
<td>30.66</td>
</tr>
<tr>
<td>Non-projection</td>
<td>32.76</td>
<td>30.38</td>
</tr>
<tr>
<td>Linear Projection</td>
<td>34.62</td>
<td>30.77</td>
</tr>
<tr>
<td>Non-linear Projection</td>
<td>34.53</td>
<td><strong>31.22</strong></td>
</tr>
</tbody>
</table>

## Analysis

1. Compositional semantics can improve translation quality
2. It is necessary to learn bilingual phrase representation
### Results

<table>
<thead>
<tr>
<th>Systems</th>
<th>Development Set</th>
<th>Test Set</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NIST 05</td>
<td>NIST 06</td>
<td>NIST 08</td>
</tr>
<tr>
<td>Baseline</td>
<td>33.53</td>
<td>30.66</td>
<td>23.32</td>
</tr>
<tr>
<td>Non-projection</td>
<td>32.76</td>
<td>30.38</td>
<td>23.12</td>
</tr>
<tr>
<td>Linear Projection</td>
<td>34.62</td>
<td>30.77</td>
<td>23.56</td>
</tr>
<tr>
<td>Non-linear Projection</td>
<td>34.53</td>
<td><strong>31.22</strong></td>
<td><strong>23.74</strong></td>
</tr>
</tbody>
</table>

### Analysis

1. Compositional semantics can improve translation quality
2. It is necessary to learn bilingual phrase representation
3. Non-linear projection is more effective than linear projection
We presented a flexible framework to learn bilingual distributed compositional representations for SMT.
Conclusions

1. We presented a flexible framework to learn bilingual distributed compositional representations for SMT.

2. Our experiments validate that Non-linear projection is more powerful than linear projection between source and target language semantic space.
Using deep learning to learn word and compositional representations
Future Work

1. Using deep learning to learn word and compositional representations
2. Finding better projection models
Future Work

1. Using deep learning to learn word and compositional representations
2. Finding better projection models
3. Speeding up decoder
Thank You!