Improving Chinese-English PropBank Alignment

Shumin Wu, Martha Palmer
University of Colorado Boulder,
Hitachi Data Systems
June 4, 2015
PropBank semantic annotation

- A corpus annotated with verbal (and more recently nominal and adjective) propositions and their arguments.
- Adds semantic information (semantic roles) to the phrase structures.
  - e.g. John gave the book to Mary
- Has a set of core (numbered) arguments and adjunct argument types
  - ARG0 is typically agent, ARG1 is typically patient or theme
Why use cross-lingual PropBank alignment?

[Diagram showing the alignment of arguments across languages.]

**Potential abstracts away syntactic differences**

- Subject-object-verb construction
- Subject-verb-object construction
Framework for PropBank alignment

1. PropBank SRL
2. Word alignment
3. Argument alignment
More PropBank alignment example

Many-to-many argument alignment
Different argument labels can align to each other
### Alignment by argument label type

<table>
<thead>
<tr>
<th>label</th>
<th>A0</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0</td>
<td><strong>1610</strong></td>
<td>79</td>
<td>25</td>
<td>-</td>
<td>-</td>
<td>9</td>
</tr>
<tr>
<td>A1</td>
<td>432</td>
<td><strong>2665</strong></td>
<td>128</td>
<td><strong>11</strong></td>
<td>-</td>
<td>142</td>
</tr>
<tr>
<td>A2</td>
<td>43</td>
<td>310</td>
<td><strong>140</strong></td>
<td>8</td>
<td>3</td>
<td>67</td>
</tr>
<tr>
<td>A3</td>
<td>2</td>
<td>14</td>
<td>21</td>
<td><strong>7</strong></td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>A4</td>
<td>1</td>
<td>37</td>
<td>9</td>
<td>3</td>
<td><strong>6</strong></td>
<td>4</td>
</tr>
<tr>
<td>V</td>
<td>25</td>
<td>28</td>
<td>22</td>
<td>1</td>
<td>-</td>
<td><strong>3278</strong></td>
</tr>
</tbody>
</table>

Chinese argument type (column) to English argument type (row) alignment on Triple-Gold Xinhua
## Alignment by argument type

<table>
<thead>
<tr>
<th>type</th>
<th>A0</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>ADV</th>
<th>BNF</th>
<th>DIR</th>
<th>DIS</th>
<th>EXT</th>
<th>LOC</th>
<th>MNR</th>
<th>PRP</th>
<th>TMP</th>
<th>TPC</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0</td>
<td>1610</td>
<td>79</td>
<td>25</td>
<td>28</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>1</td>
<td>11</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>432</td>
<td>2665</td>
<td>128</td>
<td>11</td>
<td>83</td>
<td>9</td>
<td>12</td>
<td>29</td>
<td>12</td>
<td>5</td>
<td>21</td>
<td>3</td>
<td>142</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>43</td>
<td>310</td>
<td>140</td>
<td>8</td>
<td>3</td>
<td>55</td>
<td>6</td>
<td>9</td>
<td>2</td>
<td>20</td>
<td>10</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>2</td>
<td>14</td>
<td>21</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>1</td>
<td>37</td>
<td>9</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADV</td>
<td>33</td>
<td>36</td>
<td>9</td>
<td>6</td>
<td>307</td>
<td>2</td>
<td>5</td>
<td>6</td>
<td>44</td>
<td>121</td>
<td>6</td>
<td>11</td>
<td>2</td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAU</td>
<td>1</td>
<td>1</td>
<td>16</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIR</td>
<td>1</td>
<td>13</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIS</td>
<td>2</td>
<td>69</td>
<td>40</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXT</td>
<td>4</td>
<td>26</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOC</td>
<td>23</td>
<td>65</td>
<td>13</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>162</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MNR</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>260</td>
<td>1</td>
<td>3</td>
<td>34</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOD</td>
<td>1</td>
<td>159</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEG</td>
<td>24</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PNC</td>
<td>3</td>
<td>23</td>
<td>11</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>35</td>
<td>2</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRD</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TMP</td>
<td>14</td>
<td>21</td>
<td>2</td>
<td>235</td>
<td>3</td>
<td>1</td>
<td>8</td>
<td>16</td>
<td>647</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>25</td>
<td>28</td>
<td>22</td>
<td>1</td>
<td>211</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>12</td>
<td>3278</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Chinese argument type (column) to English argument type alignment on Triple-Gold Xinhua
Many-to-many argument alignment

Objective:
choose source/target argument set to maximize total number of words in
the set while minimizing number of unaligned words between arguments

100% precision, low recall
Many-to-many argument alignment

- **Predicate-argument alignment score**
  - $a_i$: argument, $A_i$: set of arguments
  - $W_i$: set of words in $a_i$
  - $map(a_i)$: word mapping (alignment) to set of words in target language
  - Source alignment precision/recall:
    
    \[
    P_{Ic} = \frac{|(\bigcup_{i \in I} map_e(a_{i, c})) \cap (\bigcup_{j \in J} W_{j, e})|}{|\bigcup_{i \in I} map_e(a_{i, c})|}, \quad R_{Ic} = \frac{\sum_{i \in I} |W_{i, c}|}{\sum_{\forall i} |W_{i, c}|}
    \]

  - Choosing the argument set(s) to maximize alignment score:
    
    \[
    A_{I, c} = \arg \max_I \frac{2 \cdot P_{I, c} \cdot R_{I, c}}{P_{I, c} + R_{I, c}} = F_{I, c}
    \]

    \[
    A_{I, c}, A_{J, e} = \arg \max_{I, J} \frac{2 \cdot F_{I, c} \cdot F_{J, e}}{F_{I, c} + F_{J, e}} = F_{I, J}
    \]

  - Harmonic mean of precision and recall
  - F measure of source and target argument mapping score

  portion of words in aligned argument set over words in argument set

  portion of word aligned words over all words in argument set

  F measure of source and target argument mapping score
Construction of the main passage has activated in the flow of materials, the flow of people and the flow of information in the great southwest, and has promoted development in the coastal port cities’ economies.
Aligning multiple predicate-arguments

- Find the best one-to-one alignment (linear assignment problem) using Kuhn-Munkres method:

\[ g^* = \arg \max_x \sum_{i \in C} \sum_{j \in E} F_{ij} \cdot x_{ij} \quad x_{ij} \in \{0,1\}, \sum_i x_{ij} \leq 1, \sum_j x_{ij} \leq 1 \]

- Argument alignment score:

<table>
<thead>
<tr>
<th></th>
<th>activate.01</th>
<th>promote.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>搞活.01 (invigorate)</td>
<td>0.49</td>
<td>0.25</td>
</tr>
<tr>
<td>促进.01 (promote)</td>
<td>0.23</td>
<td>0.77</td>
</tr>
</tbody>
</table>

- Alignments:
  搞活.01 ↔ activate.01  促进.01 ↔ promote.01
PropBank alignment results

Semantic alignment results on broadcast conversation (bc) & Xinhua News

Are there issues? Can we improve these?
Forced alignment between verb predicates

The ancient Grand Canal ... resulting in the formation of the Grand Canal Economic Belt

Alignment of non-synonymous predicates, and predicate to core arguments
The ancient Grand Canal … resulting in the formation of the Grand Canal Economic Belt.

Using nominal predicates can produce more semantically similar alignments.
Vertically aligned table showing alignment counts with nominal predicates.

<table>
<thead>
<tr>
<th>pred. type</th>
<th>$V_c - V_e$</th>
<th>$N_c - V_e$</th>
<th>$V_c - N_e$</th>
<th>$N_c - N_e$</th>
<th>total</th>
<th>Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb only</td>
<td>4879</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4879</td>
<td>-</td>
</tr>
<tr>
<td>+Ch nom.</td>
<td>4762</td>
<td>274</td>
<td>-</td>
<td>-</td>
<td>5036</td>
<td>+3.2%</td>
</tr>
<tr>
<td>+En nom.</td>
<td>4849</td>
<td>-</td>
<td>384</td>
<td>-</td>
<td>5233</td>
<td>+7.3%</td>
</tr>
<tr>
<td>all pred.</td>
<td>4759</td>
<td>239</td>
<td>314</td>
<td>760</td>
<td>6072</td>
<td>+24.5%</td>
</tr>
</tbody>
</table>

Predicate-argument mappings counts on Xinhua News with inclusion of nominal predicates

Using nominal predicates also increases alignment coverage.
The Jiangsu Steel ... Using its own funds, it also built a large bridge spanning the canal.

<table>
<thead>
<tr>
<th></th>
<th>use.01</th>
<th>build.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>自筹.01 (self provide)</td>
<td>0.50</td>
<td>0.60</td>
</tr>
<tr>
<td>建设.01 (construct)</td>
<td>0.22</td>
<td>0.50</td>
</tr>
</tbody>
</table>

But 自筹⇔build, Arg1⇔AM-MNR are very unlikely!!
Building alignment probability

- Collect predicate-to-predicate & argument-to-argument mapping counts from corpus to build:

\[ p(pred_{j,e} | pred_{i,c}) \]

\[ p(a_{l,e} | a_{k,c}, pred_{i,c}, pred_{j,e}) \]

- Counts can be really sparse, so apply smoothing
  - Simple Good-Turing smoothing for predicate-to-predicate probability

\[ p_0 = \frac{N_1}{N} \]

\[ p_c = (c + 1) \frac{N_{c+1}}{N_c N} \]
Building alignment probability (cont)

- Smooth argument-to-argument probability with absolute discounting

\[
p(a_{lE}|a_{kC},\text{pred}_{iC},\text{pred}_{jE}) = \frac{\text{max}(\text{freq}(a_{lE}|a_{kC},\text{pred}_{iC},\text{pred}_{jE}) - d, 0)}{\sum_l(a_{lE}|a_{kC},\text{pred}_{iC},\text{pred}_{jE})} + (1 - \lambda) \cdot p_{\text{backoff}}(a_{lE})
\]

\[
p_{\text{backoff}}(a_{lE}) = p(a_{lE}|a_{kC},\text{pred}_{iC}) \quad \text{or} \quad p(a_{lE}|a_{kC},\text{pred}_{jE})
\]

- Intermediary back-off probabilities can also be smoothed:

\[
p(a_{lE}|a_{kC},\text{pred}_{iC}) = \frac{\text{max}(\text{freq}(a_{lE}|a_{kC},\text{pred}_{iC}) - d, 0)}{\sum_l(a_{lE}|a_{kC},\text{pred}_{iC})} + (1 - \lambda) \cdot p(a_{lE}|a_{kC})
\]
Alignment w/ probability model

- Update previously defined argument precision & recall
  \[ P_{kl} = (1 - \beta + \beta \cdot w(a_{l,e} \mid \text{pred}_{i,c}, \text{pred}_{j,e}, a_{k,c})) P_{kl} \]
  \[ R_{kl} = (1 - \beta + \beta \cdot w(a_{k,c} \mid \text{pred}_{i,c}, \text{pred}_{j,e}, a_{l,e})) R_{kl} \]
  where \( w(a_k) = \frac{p(a_k)}{\sum_k p(a_k) \cdot p(a_k)} \), \( P_k = \frac{|\text{map}_e(a_{k,c}) \cup W_{l,e}|}{|\text{map}_e(a_{k,c})|} \), \( R_k = \frac{|\text{map}_e(a_{l,e}) \cup W_{k,c}|}{|\text{map}_e(a_{l,e})|} \).

- Update predicate-to-predicate score:
  \[ F'_{i,c} = (1 - \alpha + \alpha \cdot w(\text{pred}_{j,e} \mid \text{pred}_{i,c})) F_{i,c} \]
  \[ F'_{j,e} = (1 - \alpha + \alpha \cdot w(\text{pred}_{i,c} \mid \text{pred}_{j,e})) F_{j,e} \]

- Need to find “good” \( \alpha \) and \( \beta \) values that balances the word alignment score and predicate/argument alignment probabilities

Portion (\( \beta \)) of alignment score weighted by argument label alignment probability

Portion (\( \alpha \)) of alignment score weighted by predicate alignment probability
Alignment w/ probability model (cont)

- Find the best $\alpha$ and $\beta$ using expectation maximization (EM)
  1. Perform semantic alignment on corpus
  2. Build probability models from mapping (E step)
  3. Find $\alpha$ and $\beta$ that maximizes the sum of the mapping score of all predicate pairs in the corpus using grid search (M step)
  4. Repeat steps 2-3 until the score sum plateaus

- Data
  - 1.6M parallel sentence pairs from multiple LDC corpora
  - Stanford Chinese word segmenter, Berkeley parser, ClearSRL

- Need only 2 iterations to converge as the mapping output doesn’t change drastically

- Optimal $\alpha=0.15$, $\beta=0.1$
  - $\alpha$ value (predicate-to-predicate probability) has a larger affect on score sum than $\beta$ value (argument-to-argument probability)
The Jiangsu Steel ... Using its own funds, it also built a large bridge spanning the canal

<table>
<thead>
<tr>
<th></th>
<th>use.01</th>
<th>build.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>自筹.01 (self provide)</td>
<td>0.65</td>
<td>0.40</td>
</tr>
<tr>
<td>建设.01 (construct)</td>
<td>0.10</td>
<td>0.85</td>
</tr>
</tbody>
</table>

自筹⇔use, Arg1⇔Arg1 are much more likely
The Jiangsu Steel … Using its own funds, it also built a large bridge spanning the canal.

Alignment w/ probability example (cont)

<table>
<thead>
<tr>
<th></th>
<th>use.01</th>
<th>build.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>自筹.01 (self provide)</td>
<td>0.65</td>
<td>0.40</td>
</tr>
<tr>
<td>建设.01 (construct)</td>
<td>0.10</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Also allowed correct 建设 ⇔ build mapping
Probability model improvements

- About 1 F point predicate mapping improvement on automatic SRL&WA input

Semantic alignment results on broadcast conversation (BC) & Xinhua News
Summary

- PropBank alignment with nominal predicates
  - Improves alignment quality and coverage
- Predicate-argument alignment with probability models
  - Generate predicate-to-predicate and argument-to-argument alignment probabilities on large corpora
  - Iteratively improve model with EM
  - Provides ~1 F-point improvement w/ automatic SRL input
Future work

- Build alignment probability models based on verb classes

\[ p(pred_{j,c} | pred_{V_i,e}) \]

- Use alignment probability model to perform joint Chinese, English SRL
  - Can improve both SRL accuracy and alignment model

- PropBank alignment of other language pairs
  - Arabic-English
Questions?