Enriching Word Alignment with Linguistic Tags

*Linguistic Data Consortium, IBM*

Xuansong Li, Niyu Ge, Stephen Grimes, Stephanie M. Strassel, Kazuaki Maeda

{xuansong, sgrimes, strassel, maeda} @ldc.upenn.edu

niyuge@us.ibm.com
Outline

- Motivations
- Approaches and methodologies
- Linguistic tags
- Inter-annotator agreement
- Conclusions
Motivations

- To improve automatic word alignment quality
- To reduce data amount needed for statistic models
- Supervised models outperform traditional models
- A part of GALE by DARPA: manually aligned and tagged data – Chinese-English WA
Unified Annotation Scheme

alignment framework

minimum translation units

minimum match approach

tagging framework

linguistic tags

attachment approach
Minimum translation units: atomic

- **One to One**
  - I buy fresh flowers.
  - 我买鲜花。

- **Many to One**
  - 春节快乐
  - Happy Chinese New Year

- **Many to Many**
Attachment Approach -- for unaligned words

Attach phrase-level unaligned words

He brought the books

Unattach sentence-level/discourse-level unaligned words

We didn’t want to hurt him
Goal: tackle insertion/deletion problems

Methodologies: using linguistic tags

- Tag aligned links
  - Context-free links (2)
  - Context-dependent links (3)
  - Specific-feature links: Chinese-DE

- Tag unaligned words
  - Tags for attached words (12 types)
  - Tags for unattached words (2 types)
Context-free Links

在 学 校

at school

Semantic Links

屹立于太行山

standing tall on Taihang Mountain
Context-dependent Links

grammatically inferred link

欢迎 收看 CCTV

Welcome to CCTV

turn this success into…

把这项成果变成…

contextually inferred link
Specific Links: 的(DE)

经历过战争的人

those who have experienced wars

新技术的实质

the essence of the new technology

将军的高度警惕

great attention from the general
<table>
<thead>
<tr>
<th>Aligned &amp; Unaligned</th>
<th>Word Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Omni-func-preposition</td>
<td>Tense/Passive</td>
</tr>
<tr>
<td>Possessive</td>
<td>Measure word</td>
</tr>
<tr>
<td>Clause marker</td>
<td>Rhetorical</td>
</tr>
<tr>
<td>Sentence marker</td>
<td>Co-reference</td>
</tr>
<tr>
<td>Determiner</td>
<td>TO-infinitive</td>
</tr>
<tr>
<td>DE-modifier</td>
<td>Local context</td>
</tr>
<tr>
<td>Context-obligatory</td>
<td></td>
</tr>
<tr>
<td>Non-context-obligatory</td>
<td></td>
</tr>
<tr>
<td>Word Tag</td>
<td>Examples</td>
</tr>
<tr>
<td>------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Possessive</td>
<td>the head of the branch</td>
</tr>
<tr>
<td>Measure-word</td>
<td>一根 (one) 柱子 (pillar) [one pillar]</td>
</tr>
<tr>
<td>Tense/Passive</td>
<td>提交 (submit) 的报告 (report) [report submitted]</td>
</tr>
<tr>
<td>Context-obligatory</td>
<td>不 (not) 好 (easy) 掌握 (control), 凭 (by) 经验 (experience) [It is not easy to control, you do by experience]</td>
</tr>
<tr>
<td>Non-context-obligatory</td>
<td>他 (he) 都已经 (already) 签 (sign) 合同了 (contract) [He already signed a contract]</td>
</tr>
</tbody>
</table>
### Chinese-English Alignment

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Char-Count</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>NW1</td>
<td>306</td>
<td>97.3%</td>
<td>95.7%</td>
<td>96.5%</td>
</tr>
<tr>
<td>NW2</td>
<td>185</td>
<td>95.3%</td>
<td>96.2%</td>
<td>95.7%</td>
</tr>
<tr>
<td>NW3</td>
<td>365</td>
<td>90.4%</td>
<td>91.2%</td>
<td>90.8%</td>
</tr>
<tr>
<td>NW4</td>
<td>431</td>
<td>90.8%</td>
<td>92.6%</td>
<td>91.2%</td>
</tr>
</tbody>
</table>
**Inter-Annotator Agreement (2)**

### Chinese-English Tagging

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Chi. Char</th>
<th>Eng. Word</th>
<th>Link Count</th>
<th>Same Tag</th>
<th>Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>NW1</td>
<td>306</td>
<td>233</td>
<td>186</td>
<td>683</td>
<td>94.2%</td>
</tr>
<tr>
<td>NW2</td>
<td>185</td>
<td>131</td>
<td>105</td>
<td>392</td>
<td>93.1%</td>
</tr>
</tbody>
</table>
Conclusion

- Unified annotation scheme
- Manually aligned and tagged corpora at LDC
- Annotation guidelines available at: http://projects.ldc.upenn.edu/gale/task Specifications/
- Annotation toolkit available soon
- On-going project: more data in pipeline
- Acknowledgements to GALE of DARPA
Thank You!
<table>
<thead>
<tr>
<th>Genre</th>
<th>File</th>
<th>Char</th>
<th>Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newswire</td>
<td>579</td>
<td>225645</td>
<td>5015</td>
</tr>
<tr>
<td>Broadcast News</td>
<td>28</td>
<td>183400</td>
<td>6376</td>
</tr>
<tr>
<td>Broadcast</td>
<td>34</td>
<td>306497</td>
<td>12050</td>
</tr>
<tr>
<td>Conversation</td>
<td>747</td>
<td>229799</td>
<td>9382</td>
</tr>
<tr>
<td>Total</td>
<td>1388</td>
<td>945341</td>
<td>32823</td>
</tr>
</tbody>
</table>
Annotation Rate

Average skill, speed and difficulty level

- First pass alignment: 10,000w/10h
- Second pass alignment: 10,000w/6h
- First pass tagging: 10,000w/7h
- Second pass tagging: 10,000w/5h