Contrastive Unsupervised Word Alignment with Non-Local Features

Yang Liu and Maosong Sun
Word Alignment

• Word alignment: aligning words between two languages

he made a speech at the meeting
Approaches

• **Generative** [Brown et al., 1993; Vogel et al., 1996, Liang et al., 2006]
  
  • **pros**: no need for labeled data
  
  • **cons**: hard to extend

• **Discriminative** [Taskar et al., 2005; Moore et al., 2006; Liu et al., 2010]
  
  • **pros**: easy to extend
  
  • **cons**: rely on labeled data
Latent-Variable Log-Linear Models

\[
P(x; \theta) = \sum_{y \in \mathcal{Y}(x)} P(x, y; \theta) = \sum_{y \in \mathcal{Y}(x)} \exp(\theta \cdot \phi(x, y)) / Z(\theta)
\]
Challenge

training data \( \{ x^{(i)} \}_{i=1}^{I} \)

objective

\[
L(\theta) = \sum_{i=1}^{I} \log \sum_{y \in \mathcal{Y}(x^{(i)})} \exp(\theta \cdot \phi(x^{(i)}, y)) - \log Z(\theta)
\]

derivative

\[
\frac{\partial L(\theta)}{\partial \theta_k} = \sum_{i=1}^{I} \mathbb{E}_{y|x^{(i)}; \theta} [\phi_k(x^{(i)}, y)] - \mathbb{E}_{x,y; \theta} [\phi_k(x, y)]
\]

intractable to calculate two feature expectations

[Smith and Eisner, 2005; Berg-Kirkpatrick et al., 2010; Dyer et al., 2011]
Idea

observation

he made a speech at the meeting

noise

talk a meeting she at the made

Intuition: observations have higher probabilities than noises
Contrastive Learning

Training data: \( \{ \langle x^{(i)}, \tilde{x}^{(i)} \rangle \}_{i=1}^I \)

Objective:

\[
J(\theta) = \log \prod_{i=1}^I \frac{P(x^{(i)}; \theta)}{P(\tilde{x}^{(i)}; \theta)}
\]

Derivative:

\[
\frac{\partial J(\theta)}{\partial \theta_k} = \sum_{i=1}^I \mathbb{E}_{y|x^{(i)}; \theta} [\phi_k(x^{(i)}, y)] - \mathbb{E}_{y|\tilde{x}^{(i)}; \theta} [\phi_k(\tilde{x}^{(i)}, y)]
\]

Partition function canceled out
Concentration

- Alignments with higher probabilities are more important in calculating expectations.
Top-n Sampling

\[ \mathbb{E}_{y|x;\theta} [\phi_k(x, y)] \approx \frac{\sum_{y \in \mathcal{N}(x;\theta)} \exp(\theta \cdot \phi(x, y)) \phi_k(x, y)}{\sum_{y' \in \mathcal{N}(x;\theta)} \exp(\theta \cdot \phi(x, y'))} \]
Comparison with Gibbs Sampling

<table>
<thead>
<tr>
<th># samples</th>
<th>Gibbs</th>
<th>Top-n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.5411</td>
<td>0.1653</td>
</tr>
<tr>
<td>5</td>
<td>0.7410</td>
<td>0.1477</td>
</tr>
<tr>
<td>10</td>
<td>0.6550</td>
<td>0.1396</td>
</tr>
<tr>
<td>50</td>
<td>0.5498</td>
<td>0.1108</td>
</tr>
<tr>
<td>100</td>
<td>0.5396</td>
<td>0.1086</td>
</tr>
<tr>
<td>500</td>
<td>0.5180</td>
<td>0.0932</td>
</tr>
</tbody>
</table>

Comparison with Gibbs sampling in terms of average approximation error
## Effect of Noise Generation

<table>
<thead>
<tr>
<th>noise generation</th>
<th>French-English</th>
<th>Chinese-English</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHUFFLE</td>
<td>8.93</td>
<td>21.05</td>
</tr>
<tr>
<td>DELETE</td>
<td>9.03</td>
<td>21.49</td>
</tr>
<tr>
<td>INSERT</td>
<td>12.87</td>
<td>24.87</td>
</tr>
<tr>
<td>REPLACE</td>
<td>13.13</td>
<td>25.59</td>
</tr>
</tbody>
</table>

Effect of noise generation in terms of alignment error rate
## Final Result

<table>
<thead>
<tr>
<th>system</th>
<th>model</th>
<th>supervision</th>
<th>algorithm</th>
<th>French-English</th>
<th>Chinese-English</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIZA++</td>
<td>IBM model 4</td>
<td>unsupervised</td>
<td>EM</td>
<td>6.36</td>
<td>21.92</td>
</tr>
<tr>
<td>Berkeley</td>
<td>joint HMM</td>
<td>unsupervised</td>
<td>EM</td>
<td>5.34</td>
<td>21.67</td>
</tr>
<tr>
<td>fast_align</td>
<td>log-linear model</td>
<td>unsupervised</td>
<td>EM</td>
<td>15.20</td>
<td>28.44</td>
</tr>
<tr>
<td>Vigne</td>
<td>linear model</td>
<td>supervised</td>
<td>MERT</td>
<td>4.28</td>
<td>19.37</td>
</tr>
<tr>
<td>this work</td>
<td>log-linear model</td>
<td>unsupervised</td>
<td>SGD</td>
<td>5.01</td>
<td>20.24</td>
</tr>
</tbody>
</table>

Comparison with state-of-the-art aligners
• Word alignment is important for multilingual NLP tasks

• Unsupervised learning of latent-variable log-linear models combines the merits of generative and discriminative approaches

• We have proposed an efficient and accurate learning algorithm for unsupervised word alignment with arbitrary features

• We will apply our approach to other NLP tasks
Source code and data sets are freely available at:
http://nlp.csai.tsinghua.edu.cn/~ly/systems/TsinghuaAligner/TsinghuaAligner.html