Improving Word Alignment Using Syntactic Dependencies

Tobias Elßner

Hauptseminar Hybrid Machine Translation

12.07.2018

*Based on Ma et al. (2008)
Overview

- *Generative Word Alignment* (Brown et al., 1993)
  - Powerful modeling of word alignment
  - Difficult to include new features

- *Discriminative Word Alignment* (Moore, 2005)
  - Easy to incorporate features
  - Feature selection is crucial to the performance
Overview: Syntactic Features

- POS (Liu et al., 2005)
- Shallow Parses (Ren et al., 2007)
- Phrase & Dependency Structures (Lopez and Resnik, 2005)
- Heuristic based Models (Ayan et al., 2004)
- Inversion Transduction Grammars (Wu, 1997)
(i) Find reliable *anchor* links (→ solid line)

(ii) Align unaligned words based on syntactic dependencies (→ dashed line)
Problem Formulation

Looking for the optimal alignment $\hat{A}$ s.t.

$$\hat{A} = \arg \max_{A} p(A \mid c^{J}, e^{I}),$$

where

- $c^{J}$ is the source sentence,
- $e^{I}$ the target sentence,
- $A \subseteq \{(i, j) \mid j = 1, \ldots, J; i = 1, \ldots, I\}$
The model

\[ p(A \mid c^J, e^l) = \prod_{j=0}^{J} p(a_j \mid c^J, e^l, a^{j-1}) \]

is decomposed into two parts:

1. \( \frac{1}{Z} \cdot p_{\epsilon}(A_{\Delta} \mid c^J, e^l) \)
   
   \( \text{Anchor Alignment Model} \)

2. \( \prod_{j \in \tilde{\Delta}} p(a_i \mid c^J, e^l, a^{j-1}, A_{\Delta}) \)
   
   \( \text{Syntax-Enhanced Model} \)
Anchor Alignment Model

The anchor alignment model

\[
\frac{1}{Z} \cdot p_\epsilon (A_\Delta \mid c^J, e^I)
\]

- detects a subset \( A_\Delta \) of highly probable alignments

- Link probabilities are constrained s.t.
  \[\forall i' \neq i \in \{1, \ldots, I\} : \frac{p(j, i)}{p(j, i')} > \epsilon_1\] between
  \[\Rightarrow \text{ A link from } c_j \text{ to } e_i \text{ is the only probable one}\]

  \[\forall j' \neq j \in \{1, \ldots, J\} : \frac{p(j, i)}{p(j', i)} > \epsilon_2\]
  \[\Rightarrow \text{ A link to } c_j \text{ from } e_i \text{ is the only probable one}\]

- Additional usage of asymmetric IBM models for bidirectional word alignment
Syntactic Enhancement

The second part of the model is

\[
\prod_{j \in \tilde{\Delta}} p(a_j \mid c^j, e^l, a^{j-1}, A_\Delta),
\]

where

- \( \tilde{\Delta} \) is the set of unaligned source words
- \( a^{j-1} \) are the alignments assigned up to the \( j \) – \( th \) word

⇒ The alignments are conditioned on
  - both sentence pairs,
  - the previously aligned non-aligned words, and
  - the set of anchor links
Statistical Features: IBM Model I Score

IBM Model I is defined as

$$p(c^J, a^J \mid e^I) = \frac{p(J|I)}{(I+1)^J} \cdot \prod_{j=1}^{J} p(c_j \mid e_{a_j})$$

- $e_{a_j}$ is the English word aligned by the $j$-th Chinese word $c_j$
- $\Rightarrow$ Assymetric generative Model
- $\Rightarrow$ Computes the joint probability of a Chinese source sentence and an alignment given the target sentence
Statistical Features: Log-Likelihood Ratio

\[
\begin{array}{c|c|c}
\hline
\text{c}_j & \neg \text{c}_j \\
\hline
\text{e}_i & a & b \\
\neg \text{e}_i & c & d \\
\hline
\end{array}
\]

Contingency Table

- Maximum likelihood estimations of probabilities:
  - \( p_1 = \frac{a}{a+b} \)
  - \( p_2 = \frac{c}{c+d} \)
  - \( p = \frac{a+c}{a+b+c+d} \)

- \( B(k|n, p) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \)

- \( G^2(c_j, e_i) = -2 \log \left( \frac{B(a|a+b, p_1) B(c|c+d, p_2)}{B(a|a+b, p) B(c|c+d, p)} \right) \)

\[\Rightarrow\] Measures the informativity of an alignment from \( c_j \) to \( e_i \)
Statistical Features: POS Translation Probability

- POS-Tags provide a remedy for the data sparseness problem

\[ p(T_c \mid T_e) = \frac{\text{COL}(T_c, T_e)}{\text{COF}(T_e)} \]

- \( \text{COL}(T_c, T_e) \) is the co-occurrence of a Chinese POS-Tag \( T_c \) and an English POS-Tag \( T_e \)

- \( \text{COF}(T_e) \) is the frequency of an English POS-Tag \( T_e \)

\( \Rightarrow \) Allows maximum likelihood estimates of POS correspondences in alignments from annotated corpora
Idea

(i) Find reliable anchor links (→ solid line)

(ii) Align unaligned words based on syntactic dependencies (—→ dashed line)
Syntactic Features: Agreement I

Given a candidate link \((j, i)\),

\[
h_{DA-1} = \begin{cases} 
1 & \exists \langle c_j, R_c, c_{j'} \rangle, \langle e_i, R_e, e_{i'} \rangle \text{ with } (j', i') \in A_\Delta, \\
0 & \text{otherwise.}
\end{cases}
\]

- \(\langle c_j, R_c, c_{j'} \rangle\) denotes the dependency relation between \(c_j\) and \(c_{j'}\)
- \(\langle e_i, R_e, e_{i'} \rangle\) denotes the dependency relation between \(e_i\) and \(e_{i'}\)
Syntactic Features: Agreement II

Given a candidate link \((j, i)\),

\[
h_{DA-2} = \begin{cases} 
1 & \exists \langle c_{j'}, R_c, c_j \rangle, \langle e_{i'}, R_e, e_i \rangle \text{ with } (j', i') \in A_{\Delta}, \\
0 & \text{otherwise.}
\end{cases}
\]

- \(\langle c_{j'}, R_c, c_j \rangle\) denotes the dependency relation between \(c_{j'}\) and \(c_j\)
- \(\langle e_{i'}, R_e, e_i \rangle\) denotes the dependency relation between \(e_{i'}\) and \(e_i\)

\(h_{DA-1} / h_{DA-2}\) check if two unaligned words have a dependency relation to two anchor-aligned words
Syntactic Features: Agreement III

Given a candidate link \((j, i)\),

\[
h^{\text{DLA-1}} = \begin{cases} 
1 & \exists \langle c_j, R_c, c_{j'} \rangle, \langle e_i, R_e, e_{i'} \rangle \text{ with } (j', i') \in A_\Delta, \text{ and } R_c = R_e, \\
0 & \text{otherwise}.
\end{cases}
\]

- \(\langle c_j, R_c, c_{j'} \rangle\) denotes the dependency relation between \(c_j\) and \(c_{j'}\);
- \(\langle e_i, R_e, e_{i'} \rangle\) denotes the dependency relation between \(e_i\) and \(e_{i'}\);
- \(h^{\text{DLA-2}}\) is defined analogously.

\(\Rightarrow h^{\text{DA-1}} / h^{\text{DA-2}}\) check if two unaligned words share the same dependency relation to two anchor-aligned words.
Syntactic Features: Source/Target Word Dependencies

Given a candidate link \((j, i)\),

\[
h_{SRC-1-Label} = \begin{cases} 
1 & \exists \langle c_j, R_c, c_{j'} \rangle, \text{ with } R_c = \text{Label} \\
0 & \text{otherwise.}
\end{cases}
\]

- \(\langle c_j, R_c, c_{j'} \rangle\) denotes the dependency relation between \(c_j\) and \(c_{j'}\)

- Analogously: \(h_{SRC-2-Label}/h_{TRGT-1-Label}/h_{TRGT-2-Label}\)
  - Check for certain dependency relations with \(\text{Label}\) between an anchor-linked word \(c_j\) and a non-aligned word \(c_{j'}\)
Syntactic Features: Target Anchors

Given a candidate link \((j, i)\),

\[
h_{SRC-1-Label} = \begin{cases} 
1 & i \in a_{\Delta} \\ 
0 & \text{otherwise.}
\end{cases}
\]

⇒ Checks if a target word is already an anchor word
Relative Distortion Feature

- Idea: Use the set of anchor alignments to approximate the relative distortion between the source and target word.

- The distortion $D(l)$ is calculated for each alignment $l = (j, i)$:
  
  $D(l) = \min(|d_L|, |d_R|)$,
  
  $d_L = (j - j_L) - (i - i_L)$, $d_R = (j - j_R) - (i - i_R)$

  where $j_R, i_R$ and $j_L, i_L$ are the right- and leftmost anchor links for positions $j$ and $i$.

- The distances where divided into 4 categories:
  
  - $p(D = 0)$
  
  - $p(D = 1, D = 2)$
  
  - $p(D = 3, D = 4)$
  
  - $p(D > 4)$
To generally find an alignment $a_j$, a discriminative model is used:

$$
\hat{a}_j = \arg \max_{a_j} \left\{ p_{\lambda_1}^M (a_j \mid c^j, e^l, a^{j-1}, A_\Delta) \right\} \\
= \arg \max_{a_j} \left\{ \sum_{m=1}^{M} \lambda_m h_m (c^j, e^l, a^l, A_\Delta, T_c, T_e) \right\}
$$

- $h_m$ are the syntactic features
- $A_\Delta$ is the obtained subset of *anchor* links
- $T_c$ and $T_e$ is the dependency structure for both languages
Experimental Setup

- Bilingual *Basic Travel Expression Corpus*
  - Basic text book phrases
  - Tagged with MXPOST (Ratnaparkhi, 1996)
  - Parsed with Malt Parser (Nivre et al., 2007)

- Manual alignment of 502 sentence pairs
  - Annotated as *Sure* (*S*) and *Possible* (*P*)
  - 300 for training
  - 50 for validation
  - 152 for testing

- Alignment training with SVM

- Phrase-Based SMT: GIZA++
  - 40,458 sentence pairs in total
  - Test set: 489 sentence pairs
Three metrics to measure alignment quality:

- **Precision:** \( \frac{|A \cap P|}{|A|} \)
  - Ratio of correctly predicted alignments to all predicted alignments

- **Recall:** \( \frac{|A \cap P|}{|P|} \)
  - Ratio of correctly predicted alignments to all correct alignments

- **Alignment Error Rate:** \( AER(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \)

- For SMT-evaluation: **Bleu**
## Results: Alignment

<table>
<thead>
<tr>
<th>model</th>
<th>precision</th>
<th>recall</th>
<th>f-score</th>
<th>AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM refined</td>
<td>0.8043</td>
<td>0.7592</td>
<td>0.7811</td>
<td>0.2059</td>
</tr>
<tr>
<td>Syntax-HMM</td>
<td>0.8744</td>
<td>0.7304</td>
<td>0.7959</td>
<td>0.1845</td>
</tr>
<tr>
<td>Model 4 refined</td>
<td>0.7941</td>
<td>0.7987</td>
<td>0.7964</td>
<td>0.1929</td>
</tr>
<tr>
<td>Syntax-Model 4</td>
<td>0.8566</td>
<td>0.7685</td>
<td>0.8102</td>
<td><strong>0.1730</strong></td>
</tr>
</tbody>
</table>

Comparison between the Syntax-Enhanced and the Generative Approach
# Results: Influence on MT

<table>
<thead>
<tr>
<th>Model</th>
<th>dev. set</th>
<th>test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.5412</td>
<td>0.3510 (BP=0.96)</td>
</tr>
<tr>
<td>Syntax-HMM</td>
<td>0.6015</td>
<td>0.3409 (BP=0.86)</td>
</tr>
<tr>
<td>Syntax-Model 4</td>
<td>0.5834</td>
<td>0.3585 (BP=0.91)</td>
</tr>
</tbody>
</table>

Influence of the Model on MT
### Discussion: Weight Contribution

<table>
<thead>
<tr>
<th>Feature</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 Score</td>
<td>0.1416</td>
</tr>
<tr>
<td>POS</td>
<td>0.0540</td>
</tr>
<tr>
<td>Log-likelihood Ratio</td>
<td>0.0856</td>
</tr>
<tr>
<td>relative distortion</td>
<td>0.0606</td>
</tr>
<tr>
<td>DA-1</td>
<td>0.0227</td>
</tr>
<tr>
<td>DLA-2</td>
<td>0.0927</td>
</tr>
<tr>
<td>tgt-1-PRD</td>
<td>0.0961</td>
</tr>
<tr>
<td>tgt-2-AMOD</td>
<td>0.0621</td>
</tr>
</tbody>
</table>

**Selection of Informative Features**


