Tree-to-String Alignment Template for Statistical Machine Translation

Yang Liu, Qun Liu, and Shouxun Lin
Institute of Computing Technology
Chinese Academy of Sciences
Outline

- Introduction
- Tree-to-String Alignment Template
- Training
- Decoding
- Experiments
- Recent Advances
- Conclusion and Future Work
Current Situation of SMT

- Phrase-based models are state-of-the-art
  - Och and Ney, 2004

- Syntax-based models are in rapid development
  - Wu, 1997
  - Alshawi et al., 2000
  - Yamada and Knight, 2001
  - Melamed, 2004
  - Galley et al., 2004
  - Graehl and Knight, 2004
  - Chiang, 2005
  - Quirk et al., 2005
  - Ding and Palmer, 2005
Challenges to Syntax-based Models

- Most syntax-based models do not show improvement over phrase-based ones. Why?
  - complexity
  - non-isomorphism
  - non-perfect training data
  - ...

Our Work

- Goal: simple and powerful
- Key: tree-to-string alignment template (TAT)
- Distinctions from previous work:
  - Model the syntax of the source language
  - Exploit bilingual phrases to strengthen the TAT-based model
Outline

- Introduction
- Tree-to-String Alignment Template
- Training
- Decoding
- Experiments
- Recent Advances
- Conclusion and Future Work
Tree-to-String Alignment Template

```
NP
  NR
  NN

President Bush

NP
  NP
  LC
  CC
  NR

between United States and

NP
  DNP
  DEG
```
Translation Process: Parsing

中国 的 经济 发展

parsing

NP

DNP NP

NP DEG NN NN

NR 的 经济 发展

中国
Translation Process: Detachment

NP
  DNP  NP
  NP  DEG  NN  NN
  NR  的  经济  发展
  中国

 detachment

NP
  DNP  NP
  NP  DEG  NN  NN
  NR  的  经济  发展
  中国
Translation Process : Production

NP
DNP
NP
NP
DEG
NP
NR
China

NN
economic
NN
development

Go to page 27
Translation Process: Combination

The diagram illustrates the process of translating the Chinese phrase "经济的发展" (economic development) into English as "economic development of China." The structure shows how the components of the phrase are aligned and combined.
Log-linear Model

\[
Pr(e_1^I, z_1^K | f_1^J) = \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J, z_1^K)]}{\sum_{e_1^I, z_1^K} \exp[\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J, z_1^K)]}
\]

\[
\hat{e}_1^I = \arg\max_{e_1^I, z_1^K} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J, z_1^K) \right\}
\]
Feature Functions

\[ h_1(e_1^I, f_1^J) = \log \prod_{k=1}^{K} \frac{N(z) \cdot \delta(T(z), \tilde{T}_k)}{N(T(z))} \]

\[ h_2(e_1^I, f_1^J) = \log \prod_{k=1}^{K} \frac{N(z) \cdot \delta(T(z), \tilde{T}_k)}{N(S(z))} \]

\[ h_3(e_1^I, f_1^J) = \log \prod_{k=1}^{K} \text{lex}(T(z)|S(z)) \cdot \delta(T(z), \tilde{T}_k) \]

\[ h_4(e_1^I, f_1^J) = \log \prod_{k=1}^{K} \text{lex}(S(z)|T(z)) \cdot \delta(T(z), \tilde{T}_k) \]

TAT penalty

\[ h_5(e_1^I, f_1^J) = K \]

trigram LM

\[ h_6(e_1^I, f_1^J) = \log \prod_{i=1}^{I} p(e_i|e_{i-2}, e_{i-1}) \]

word penalty

\[ h_7(e_1^I, f_1^J) = I \]
Outline

- Introduction
- Tree-to-String Alignment Template
- Training
- Decoding
- Experiments
- Recent Advances
- Conclusion and Future Work
Extract TATs

- TSA (Tree-String-Alignment)
  - bilingual phrase with tree over the source string
- Bottom-up strategy
- Impose several restrictions to reduce the magnitude
  - maximal height of the tree
  - maximal number of children of a node
  - both the head and tail of target string must be aligned
An Example

President Bush made a speech
An Example

President Bush made a speech

布什 总统 发表 演讲

President Bush made a speech

布什

Bush
An Example

President Bush made a speech

布什 总统 发表 演讲
An Example

President Bush made a speech

布什 总统 发表 演讲

made
An Example

President Bush made a speech
An Example

President Bush made a speech
An Example

President Bush made a speech

made a speech
An Example

President Bush made a speech

h=2, c=2
Outline

- Introduction
- Tree-to-String Alignment Template
- Training
- Decoding
- Experiments
- Recent Advances
- Conclusion and Future Work
Decoding

- Bottom-up beam search
- For each sub tree, we compute a list of candidate translations (derivations)
- A candidate translation contains the following information:
  - TAT sequence
  - partial translation
  - accumulated feature values
  - accumulated probability
  - ...
Default TAT

NR  
中国

NP
DNP  NP
NP  DEG  NN  NN
NR  的  经济  发展
中国

construct default TATs

NR  
中国

NP
DNP  NP
An Example

Usable TAT

NR

中国

Derivation

( NR 中国 )  China  1:1

Translation

China
An Example

Usable TAT (default)

Derivation

| (NP (NR)) | X | 1:1 |
| (NR 中国) | China | 1:1 |

Translation

China
An Example

Usable TAT (default)

Derivation

Translation

中国的经济发展

的
An Example

Usable TAT (default)

Derivation

<table>
<thead>
<tr>
<th>(DNP (NP) (DEG))</th>
<th>X1</th>
<th>X2</th>
<th>1:1 2:2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(NP (NR 中国))</td>
<td></td>
<td></td>
<td>1:1</td>
</tr>
<tr>
<td>(DEG 的)</td>
<td></td>
<td></td>
<td>1:1</td>
</tr>
</tbody>
</table>

Translation

China 的
An Example

Usable TAT

Derivation

Translation

economic
An Example

Usable TAT

Derivation

Translation

development

development
An Example

Usable TAT

Derivation

Translation
development
An Example

Usable TAT

Derivation

Translation

economic development of China
Recombination

The economic development of China is very rapid.
The economic development of China is quite rapid.
The economic developing of Chinese is rapid.
The economic development of Chinese are quite rapid.

To perform recombination, we combine candidate translations that share the same leading and trailing bigrams (for trigram language model) in each stack.
Pseudo Code

initialize derivationStackVec[1 .. nodeCount]
for i=1 to nodeCount
    for each TAT usable to the subtree
        compute derivations
        add the derivations to derivationStackVec[i]
        prune derivationStackVec[i]
find the best derivation in derivationStackVec[nodeCount]

Note that we have not developed n-best list generation algorithm yet. To perform minimum error rate training, we just use the translations in the final stack.
Treat BPs as TATs

- **Why?**
  - bilingual phrases are “cheaper” than TATs
  - syntactic analysis is not reliable
  - lose useful non-syntactic phrase pairs due to strict restrictions

- **How?**
  - treat bilingual phrases as special TATs without tree over the source side
An Example

TAT table

NP

NR
布什
NN
总统

President Bush

0.4 0.2 0.3 0.5

BP table

布什 总统

President Bush

0.3 0.6 0.2 0.4
An Example

TAT table

NP

NR
布什

NN
总统

BP table

NP

NR
布什

NN
总统

0.4 0.2 0.3 0.5

0.3 0.6 0.2 0.4
An Example

[Diagram of a tree structure showing NPs and NRs labeled in Chinese and English, with a comparison of confidence scores]
Outline

- Introduction
- Tree-to-String Alignment Template
- Training
- Decoding
- Experiments
- Recent Advances
- Conclusion and Future Work
Experiments

- Baseline: Pharaoh (Koehn et al., 2004)
- Training corpus: 31,149 sentence pairs with 843K Chinese words and 949K English words
- Development set: 2002 NIST Chinese-to-English test set (571 of 878 sentences)
- Test set: 2005 NIST Chinese-to-English test set (1,082 sentences)
Data Process

bilingual corpus

GIZA++ and refinement

word-aligned bilingual corpus

Pharaoh training toolkits

Chinese parser

word-aligned, source-side parsed bilingual phrases

bilingual phrases

Lynx training toolkits

TATs
Some Tools

- Evaluation: mteval-v11b.pl
- Language Model: SRI Language Modeling Toolkits (Stolcke, 2002)
- Significance test: Zhang et al., 2004
- Parser: Xiong et al., 2005
- Minimum error rate training: optimizeV5IBM_BLEU.m (Venugopal and Vogel, 2005)
Results

<table>
<thead>
<tr>
<th>System</th>
<th>Features</th>
<th>BLEU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pharaoh</td>
<td>$d + \phi(e</td>
<td>f)$</td>
</tr>
<tr>
<td></td>
<td>$d + \text{lm} + \phi(e</td>
<td>f) + \text{wp}$</td>
</tr>
<tr>
<td></td>
<td>$d + \text{lm} + \phi(f</td>
<td>e) + \text{lex}(f</td>
</tr>
<tr>
<td>Lynx</td>
<td>$h_1$</td>
<td>0.1639 ± 0.0077</td>
</tr>
<tr>
<td></td>
<td>$h_1 + h_6 + h_7$</td>
<td>0.2100 ± 0.0089</td>
</tr>
<tr>
<td></td>
<td>$h_1 + h_2 + h_3 + h_4 + h_5 + h_6 + h_7$</td>
<td>0.2178 ± 0.0080</td>
</tr>
</tbody>
</table>

Comparison of Pharaoh and Lynx with different feature settings

Lynx achieves an absolute improvement of 0.9% (4.3% relative) over Pharaoh in terms of BLEU score. This difference is statistically significant ($p < 0.01$).
Effect of Using BPs

Using bilingual phrases brings an absolute improvement of 0.6% in terms of BLEU score
Outline

- Introduction
- Tree-to-String Alignment Template
- Training
- Decoding
- Experiments
- Recent Advances
- Conclusion and Future Work
Recent Advances

- Scaling to large data
- Use BPs to improve fluency
Scaling to Large Data

- Bilingual corpus (train BPs and TATs)
  - 2.6M sentence pairs (68.1M Chinese words and 73.8M English words)
  - Use all the data to obtain BPs and a portion of 800K pairs to obtain TATs

- Monolingual corpora (train LM)
  - English side of the bilingual corpus (73.8M words)
  - Xinhua portion of Gigaword corpus (181M words)
Using BPs to Improve Fluency

Problem with Lynx:

International will severely punish cheat behaviour on the football field

International Executive Committee also announces that some reform measures.

How could this happen?
Two Parse Trees

the strength of BPs is restricted!
Solution

- When the search ends, refine the output by replacing strings with more fluent ones with the help of alignment.
- Use language model to measure fluency
- If there are more than one candidates, choose the one with highest score (take translation probabilities into account)
INSTITUTE OF COMPUTING TECHNOLOGY

An Example

Executive Committee also announces that some reform measures.

international 足联 执委会 还宣布了 一些改革措施
An Example

FIFA Executive Committee also announces that some reform measures.

国际足联执委会还宣布了一些改革措施。
An Example

FIFA Executive Committee also announced some reform measures.
# Results

<table>
<thead>
<tr>
<th>Training Data (pairs)</th>
<th>Language Model</th>
<th>Improve Fluency</th>
<th>BLEU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAT</td>
<td>BP</td>
<td>Data (words)</td>
<td>Order</td>
</tr>
<tr>
<td>31K</td>
<td>-</td>
<td>949K</td>
<td>one 3-gram</td>
</tr>
<tr>
<td>31K</td>
<td>31K</td>
<td>949K</td>
<td>one 3-gram</td>
</tr>
<tr>
<td>31K</td>
<td>800K</td>
<td>73M</td>
<td>one 3-gram</td>
</tr>
<tr>
<td>800K</td>
<td>2.6M</td>
<td>73M</td>
<td>one 3-gram</td>
</tr>
<tr>
<td>800K</td>
<td>2.6M</td>
<td>73M</td>
<td>181M</td>
</tr>
<tr>
<td>800K</td>
<td>2.6M</td>
<td>73M</td>
<td>181M</td>
</tr>
<tr>
<td>800K</td>
<td>2.6M</td>
<td>73M</td>
<td>181M</td>
</tr>
</tbody>
</table>

Results of Lynx on test set with various settings.
Outline

- Introduction
- Tree-to-String Alignment Template
- Training
- Decoding
- Experiments
- Recent Advances
- Conclusion and Future Work
Conclusion

- The TAT-based translation model is simple and powerful
- Bilingual phrases can be used to strengthen the TAT-based model:
  - Treat them as special TATs
  - Use them to improve the fluency of the output
Future Work

- N-best list generation
- Better training methods for TATs
- Upgrading Tree-to-String to Forest-to-String (allow forest instead of tree over source string)
Thanks!