Parser Self-Training for Syntax-Based Machine Translation

Nara Institute of Science and Technology
Augmented Human Communication Laboratory

Makoto Morishita, Koichi Akabe, Yuto Hatakoshi, Graham Neubig, Koichiro Yoshino, Satoshi Nakamura

2015/12/03
IWSLT 2015
Background
Phrase-Based Machine Translation

[Koehn et al., 2003]

- Translate and reorder by phrases.
  - Easy to learn translation model.
  - Low translation accuracy on language pairs with different word order.

Translation Model

Reordering Model

John hit a ball

ジョンは 打球を

ジョンは ボールを 打った
Tree-to-String Machine Translation

[Liu et al., 2006]

Use the source language parse tree in translation
- High translation accuracy on language pairs with different word order.
- Translation accuracy is affected greatly by the parser accuracy.
Forest-to-String Machine Translation

[Mi et al., 2008]

- Use the source language parse forest in translation
  - Decoder can choose the parse tree that has high translation probability from the parse tree candidates

[Makoto Morishita, AHC Lab, NAIST]
Parser Self-Training [McClosky et al., 2006]

- Use the parser output as training data.
- Improve the parser accuracy.
  - Parser is adapted to the target domain.
Self-Training for Preordering

By selecting the parse trees, more effective self-training (Targeted Self-Training).
- Use only high scored parse trees.
- However, in this method, we need hand-aligned data.
- It is costly to make hand-aligned data.

[Katz-Brown et al., 2011]
Proposed Method
Proposed Method

1. **Input sentence**
   - Sentence is fed into a **Parser**.
2. **Parse forest**
   - The parser outputs a forest of parse trees.
3. **Use as training data**
   - Parse trees are used to train a **Forest-to-String Decoder**.
4. **Translated sentence and parse tree used in translation**
   - The decoder generates a translated sentence and corresponding parse tree.
5. **Evaluation using MT automatic evaluation metrics**
   - The system evaluates the translation quality using automatic metrics.

- **Targeted Self-Training using MT automatic evaluation metrics**
  - Low cost and accurate evaluation
Selection Methods

- **Parse tree selection**
  - Select a parse tree to use from a single sentence

- **Sentence selection**
  - Select the sentences to use from the entire corpus
Parse Tree Selection

- **Parser 1-best**
  - Use the parser 1-best tree.
  - Traditional self-training [McClosky et al. 2006].

- **Decoder 1-best**
  - Use the parse tree used in translation.

- **Evaluation 1-best**
  - Among the translation candidates, use the parse tree used in highest scored translation.
Decoder 1-best

- Use the parse tree used in translation.
Evaluation 1-best

- Among the translation candidate, use the parse tree used in highest scored translation.
- This highest scored translation is called Oracle translation.
Sentence Selection

- **Random**
  - Select sentences randomly from the corpus.
  - Traditional self-training.

- **Threshold** of the evaluation score
  - Use sentences that score over the threshold.

- **Gain** of the evaluation score
  - Use sentences that have a large gain in score between decoder 1-best and oracle translation.
Threshold of the Evaluation Score

Selection based on the score

Score $\geq$ Threshold

Use

Score $<$ Threshold

Do not use

High scored translation and parse tree (Oracle translation)

Threshold of the evaluation score
- Use sentences that score over threshold.
Gain of the Evaluation Score

- Use sentences that have a large gain in score between decoder 1-best and oracle translation.
Experiments
Experimental Setup (for Self-Training)

Existing model
Japanese Dependency Corpus (7k)

Parser self-training
Parallel corpus (ASPEC 2.0M)

Source language
Target language

Evaluation (BLEU+1)

Parser (Egret)

Forest-to-String Decoder (Travatar)
In these experiments, we focused on Japanese-English, Japanese-Chinese translation.
## Experiment Results
(Japanese-English Translation)

<table>
<thead>
<tr>
<th>Tree Selection</th>
<th>Sentence Selection</th>
<th>Sentences (k)</th>
<th>BLEU</th>
<th>RIBES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>-</td>
<td>-</td>
<td>23.83</td>
<td>72.27</td>
</tr>
<tr>
<td>Parser 1-best</td>
<td>Random</td>
<td>96</td>
<td>23.66</td>
<td>71.77</td>
</tr>
<tr>
<td>Decoder 1-best</td>
<td>Random</td>
<td>97</td>
<td>23.81</td>
<td>72.04</td>
</tr>
<tr>
<td>Oracle</td>
<td>Random</td>
<td>97</td>
<td>23.93</td>
<td>72.09</td>
</tr>
</tbody>
</table>
Oracle Translation Score Distribution

- It contains a lot of noisy sentences.
## Experiment Results (Japanese-English Translation)

<table>
<thead>
<tr>
<th>Tree Selection</th>
<th>Sentence Selection</th>
<th>Sentences (k)</th>
<th>BLEU</th>
<th>RIBES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>-</td>
<td>23.83</td>
<td>72.27</td>
</tr>
<tr>
<td>Parser 1-best</td>
<td>Random</td>
<td>96</td>
<td>23.66</td>
<td>71.77</td>
</tr>
<tr>
<td>Decoder 1-best</td>
<td>Random</td>
<td>97</td>
<td>23.81</td>
<td>72.04</td>
</tr>
<tr>
<td>Oracle</td>
<td>Random</td>
<td>97</td>
<td>23.93</td>
<td>72.09</td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 Threshold</td>
<td>120</td>
<td><strong>24.26</strong></td>
<td>72.38</td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 Gain</td>
<td>100</td>
<td>*24.22</td>
<td>72.32</td>
</tr>
</tbody>
</table>

* : p < 0.05  ** : p < 0.01

By self-training, the accuracy significantly improved
## Manual Evaluation

<table>
<thead>
<tr>
<th>Tree selection</th>
<th>Sentence selection</th>
<th>Score</th>
<th>Significance between Baseline</th>
<th>Significance between Parser -best</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>2.38</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Parser 1-best</td>
<td>Random</td>
<td>2.42</td>
<td>No</td>
<td>—</td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 Threshold</td>
<td>2.50</td>
<td>Yes (99% level)</td>
<td>Yes (90% level)</td>
</tr>
</tbody>
</table>

Score range is 1 to 5

- We could verify that our method is effective.
## Example of an improvement

<table>
<thead>
<tr>
<th>Source</th>
<th>C投与群ではRの活動を240分にわたって明らかに増強した</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>in the C - administered group, thermal reaction clearly increased the activity of R for 240 minutes.</td>
</tr>
<tr>
<td>Baseline</td>
<td>for 240 minutes clearly enhanced the activity of C administration group R.</td>
</tr>
<tr>
<td>Self-Trained</td>
<td>for 240 minutes clearly enhanced the activity of R in the C - administration group.</td>
</tr>
</tbody>
</table>
Before Self-Training

C  投与  群
administered  group

で  は  R  の  活動  を
in  TOP  of  activity  OBJ

AUX_SYMP
SYMP
AUX_SYMP
AUX_VP
SYMP
PP
NP
P
N
PP
P
P
N
After Self-Training

C administered group in

R の 活動 を

NP 投与 群 で は

TOP

SYM

of activity OBJ
## Experiment Results (Japanese-Chinese Translation)

<table>
<thead>
<tr>
<th>Tree Selection</th>
<th>Sentence Selection</th>
<th>Sentences (k)</th>
<th>BLEU</th>
<th>RIBES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>-</td>
<td>29.60</td>
<td>81.32</td>
</tr>
<tr>
<td>Parser 1-best</td>
<td>Random</td>
<td>129</td>
<td>29.75</td>
<td>**81.55</td>
</tr>
<tr>
<td>Decoder 1-best</td>
<td>Random</td>
<td>130</td>
<td>29.76</td>
<td>*81.53</td>
</tr>
<tr>
<td>Oracle</td>
<td>Random</td>
<td>130</td>
<td>**29.89</td>
<td>**81.66</td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 Threshold</td>
<td>82</td>
<td>*29.86</td>
<td>**81.60</td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 Gain</td>
<td>100</td>
<td>*29.85</td>
<td>**81.59</td>
</tr>
<tr>
<td>Oracle (ja-en)</td>
<td>BLEU+1 Threshold</td>
<td>120</td>
<td>*29.87</td>
<td>*81.58</td>
</tr>
</tbody>
</table>

* : p < 0.05  ** : p < 0.01

- By self-training, the accuracy significantly improved
- By using ja-en self-trained model, it also improved the accuracy.
Experiment Results (Japanese-Chinese Translation)

<table>
<thead>
<tr>
<th>Tree Selection</th>
<th>Sentence Selection</th>
<th>Sentences (k)</th>
<th>BLEU</th>
<th>RIBES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>-</td>
<td>29.60</td>
<td>81.32</td>
</tr>
<tr>
<td>Parser 1-best</td>
<td>Random</td>
<td>129</td>
<td>29.75</td>
<td>**81.55</td>
</tr>
<tr>
<td>Decoder 1-best</td>
<td>Random</td>
<td>130</td>
<td>29.76</td>
<td>* 81.53</td>
</tr>
<tr>
<td>Oracle</td>
<td>Random</td>
<td>130</td>
<td>**29.89</td>
<td>**81.66</td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 Threshold</td>
<td>82</td>
<td>* 29.86</td>
<td>**81.60</td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 Gain</td>
<td>100</td>
<td>* 29.85</td>
<td>**81.59</td>
</tr>
<tr>
<td>Oracle (ja-en)</td>
<td>BLEU+1 Threshold</td>
<td>120</td>
<td>* 29.87</td>
<td>* 81.58</td>
</tr>
</tbody>
</table>

* : p < 0.05  ** : p < 0.01

- By self-training, the accuracy significantly improved
- By using ja-en self-trained model, it also improved the accuracy.
Parser Accuracy
Experimental Setup

- 100 manually annotated trees


- We test Ja-En parsers.
### Experiment Results

<table>
<thead>
<tr>
<th>Tree selection</th>
<th>Sentence selection</th>
<th>Recall</th>
<th>Precision</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>84.88</td>
<td>84.77</td>
<td>84.83</td>
</tr>
<tr>
<td>Parser 1-best</td>
<td>Random</td>
<td>86.52</td>
<td>86.41</td>
<td>* 86.46</td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 Threshold</td>
<td>88.13</td>
<td>88.01</td>
<td>**88.07</td>
</tr>
</tbody>
</table>

* : p < 0.05  ** : p < 0.01

Our method improves not only MT results, but also parser accuracy itself.
Conclusion
Conclusion

- By our proposed self-training method, translation and parser accuracy improved.

- Self-Training does not rely on target language
  - By using Ja-En self-trained model, Ja-Zh translation accuracy improved.

- Future work
  - Verify this method is applicable in other languages.
  - Self-training using several target languages data.
  - Test the effect when performing the parser self-training repeatedly.
END
## Experiment Results (Japanese-English Translation)

<table>
<thead>
<tr>
<th>Tree Selection</th>
<th>Sentence Selection</th>
<th>Sentences (k)</th>
<th>BLEU</th>
<th>RIBES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>-</td>
<td>-</td>
<td>23.83</td>
<td>72.27</td>
</tr>
<tr>
<td>Parser 1-best</td>
<td>Random</td>
<td>96</td>
<td>23.66</td>
<td>71.77</td>
</tr>
<tr>
<td>Decoder 1-best</td>
<td>Random</td>
<td>97</td>
<td>23.81</td>
<td>72.04</td>
</tr>
<tr>
<td>Oracle</td>
<td>Random</td>
<td>97</td>
<td>23.93</td>
<td>72.09</td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 ≥ 0.7</td>
<td>206</td>
<td><strong>24.27</strong></td>
<td>72.38</td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 ≥ 0.8</td>
<td>120</td>
<td><strong>24.26</strong></td>
<td>72.38</td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 ≥ 0.9</td>
<td>58</td>
<td><strong>24.26</strong></td>
<td>72.49</td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 Gain</td>
<td>100</td>
<td>*24.22</td>
<td>72.32</td>
</tr>
</tbody>
</table>

* : p < 0.05
** : p < 0.01
## Experiment Results
(Japanese-Chinese Translation)

<table>
<thead>
<tr>
<th>Tree Selection</th>
<th>Sentence Selection</th>
<th>Sentences (k)</th>
<th>BLEU</th>
<th>RIBES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td>-</td>
<td>29.60</td>
<td>81.32</td>
</tr>
<tr>
<td>Parser 1-best</td>
<td>Random</td>
<td>129</td>
<td>29.75</td>
<td><strong>81.55</strong></td>
</tr>
<tr>
<td>Decoder 1-best</td>
<td>Random</td>
<td>130</td>
<td>29.76</td>
<td>* 81.53</td>
</tr>
<tr>
<td>Oracle</td>
<td>Random</td>
<td>130</td>
<td><strong>29.89</strong></td>
<td><strong>81.66</strong></td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 ≥ 0.7</td>
<td>240</td>
<td><strong>29.86</strong></td>
<td><strong>81.60</strong></td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 ≥ 0.8</td>
<td>150</td>
<td><strong>29.91</strong></td>
<td>81.47</td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 ≥ 0.9</td>
<td>82</td>
<td>* 29.86</td>
<td><strong>81.60</strong></td>
</tr>
<tr>
<td>Oracle</td>
<td>BLEU+1 Gain</td>
<td>100</td>
<td>* 29.85</td>
<td><strong>81.59</strong></td>
</tr>
<tr>
<td>Oracle (ja-en)</td>
<td>BLEU+1 ≥ 0.8</td>
<td>120</td>
<td>* 29.87</td>
<td>* 81.58</td>
</tr>
</tbody>
</table>

* : p < 0.05  
** : p < 0.01
Why decoder 1-best parse tree is better than parser 1-best?

- Probability considered in Forest-to-String translation
  - Parse tree probability
  - Translation model
  - Language model

- The rule that use correct tree have high probability on translation model.
  - The rule that use incorrect tree have low probability.

- By using language model,
  the correct parse tree tends to be chosen.
  - The correct tree have high probability on language model.