Learning to Parse and Translate Improves Neural Machine Translation

Akiko Eriguchi¹, Yoshimasa Tsuruoka¹, and Kyunghyun Cho²
¹ The University of Tokyo
² New York University
Overview

- Syntactic Neural Machine Translation (NMT) model
  - generates a translation
  - parses the translated sentence (optionally at test time)

“白い犬がソファの上で寝ている。”

Our Proposed NMT Model

ROOT

The white dog is sleeping on the sofa.
Neural Machine Translation

- Recurrent Neural Network-based (RNN) model
  - softly align a target word with source words with $\alpha_{ij}$
    (Bahdanau et al., 2015; Luong et al., 2015)
  - directly optimize the conditional language model $p(y|x)$

(target $y$)

$\alpha_{ij} = 0.85$

$\alpha_{ij} = 0.05$

(source $x$)
Motivation

- RNN-based NMT model captures a sequence of words and does NOT utilize syntactic information explicitly.
- To decode a translation with its parsed tree is expected to be useful (e.g. in selecting a grammatical sentence)

The white dog is sleeping on the sofa.
Recurrent Neural Network Grammars (RNNG)
(Dyer et al., 2016)

- Joint model of shift-reduce parsing and language model
- Selects an action out of three types of actions from the composed vectors of Stack, Action, and Buffer
- Achieved the state-of-the-art performance in both tasks
(Kuncoro et al., 2017)
Recurrent Neural Network Grammars (RNNG)

(Dyer et al., 2016)

- Joint model of shift-reduce parsing and language model
- Selects an action out of three types of actions from the composed vectors of Stack, Action, and Buffer
- achieved the state-of-the-art performance in both tasks

(Kuncoro, et al 2017)

Generate a word from a vocabulary $|K|$

{Shift, Reduce Right, Reduce Left}

syntactic sentence vector

sequential sentence vector

Stack  Action  Buffer

2017/08/01 ACL 2017
Recurrent Neural Network Grammars (RNNG)

(Dyer et al., 2016)

- Joint model of shift-reduce parsing and language model
  - Selects an action out of three types of actions from the composed vectors of Stack, Action, and Buffer
  - achieved the state-of-the-art performance in both tasks (Kuncoro, et al 2017)

"Reduce” the words in the Stack

{Shift, Reduce Right, Reduce Left}

syntactic sentence vector

stack

Action

Buffer

sequential sentence vector

sequential vector
Proposed Model: NMT + RNNG

- Objective function: $J(\Theta) = \sum \log p(y, a| x)$ (y's parsed actions $a$)

{Shift, Reduce Right, Reduce Left} (actions $a$)

The

Stack

Action

Buffer

Sentence vector from Bi-directional Encoder

The white dog is sleeping on the sofa.
Proposed Model: NMT + RNNG

- Objective function: $J(\Theta) = \sum \log p(y, a | x)$ (y’s parsed actions $a$)

{Shift, Reduce Right, Reduce Left} (actions $a$)

Sentence vector from Bi-directional Encoder

The white dog is sleeping on the sofa.
Proposed Model: NMT + RNNG

- Objective function: \( J(\Theta) = \sum \log p(y, a | x) \) (\( y \)'s parsed actions \( a \))

{Shift, Reduce Right, Reduce Left} (actions \( a \))

The white dog is sleeping on the sofa.

(The white dog is sleeping on the sofa.)
Proposed Model: NMT + RNNG

- Objective function: $J(\Theta) = \sum \log p(y, a|x)$ (y’s parsed actions $a$)

Sentence vector from Bi-directional Encoder

The white dog

{Shift, Reduce Right, Reduce Left} (actions $a$)

(target $y$)

(white dog is sleeping on the sofa.)
Proposed Model: NMT + RNNG

- Objective function: $J(\Theta) = \Sigma \log p(y, a | x)$ \hspace{0.5cm} (y’s parsed actions $a$)

The white dog

Sentence vector from Bi-directional Encoder

Stack

Action

Buffer

(white dog sleeping on the sofa)
Proposed Model: NMT + RNNG

- Objective function: \( J(\Theta) = \sum \log p(y, a | x) \) \( (y's \ parsed \ actions \ a) \)

Shift-Reduce Parsing Model (RNNGs)

\[
\text{Sentence vector from Bi-directional Encoder} \quad \rightarrow \quad \text{Stack} \quad \rightarrow \quad \text{Action} \quad \rightarrow \quad \text{Buffer} \quad \rightarrow \quad \text{Translation Model (Luong et al., 2015)}
\]

(The white dog is sleeping on the sofa.)
Experimental Settings

- Experiments on four translation tasks in the language pairs of {JP, RU, CZ, DE}-EN
  - JP-EN: ASPEC corpus (Nakazawa et al., 2016)
  - {RU, CZ, DE}-EN: News Commentary v8
- Parse the target sentences of the training data by SyntaxNet (Andor et al., 2016) with the dependency labels
  - At test time, SyntaxNet is not used
- 256 dimensional single-layer (Stack-)LSTM units
# Experimental Results

- Our model achieved the better accuracies than the baseline NMT model in BLEU (Papineni et al., 2002) except DE-EN translation and in RIBES (Isozaki et al., 2010)

## BLEU

<table>
<thead>
<tr>
<th></th>
<th>JP-EN</th>
<th>RU-EN</th>
<th>CS-EN</th>
<th>DE-EN</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td>17.88</td>
<td>12.03</td>
<td>11.22</td>
<td>16.61</td>
</tr>
<tr>
<td>NMT + RNNG</td>
<td><strong>18.84</strong></td>
<td><strong>12.46</strong></td>
<td><strong>12.06</strong></td>
<td>16.41</td>
</tr>
</tbody>
</table>

## RIBES

<table>
<thead>
<tr>
<th></th>
<th>JP-EN</th>
<th>RU-EN</th>
<th>CS-EN</th>
<th>DE-EN</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT</td>
<td>71.27</td>
<td>69.56</td>
<td>69.59</td>
<td>73.75</td>
</tr>
<tr>
<td>NMT + RNNG</td>
<td><strong>72.25</strong></td>
<td><strong>71.04</strong></td>
<td><strong>70.39</strong></td>
<td><strong>75.03</strong></td>
</tr>
</tbody>
</table>

※ significant difference by bootstrap resampling (p < 0.05) (Koehn, 2004)
Which component in RNNG is effective?

- We removed each component in RNNG
  - Stack-only RNNG achieved the state-of-the-art accuracy in parsing task (Kuncoro et al., 2017)

<table>
<thead>
<tr>
<th>Component</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMT + RNNG</td>
<td>18.60</td>
</tr>
<tr>
<td>w/o Buffer</td>
<td>18.02</td>
</tr>
<tr>
<td>w/o Action</td>
<td>17.94</td>
</tr>
<tr>
<td>w/o Stack</td>
<td>17.58</td>
</tr>
<tr>
<td>NMT</td>
<td>17.75</td>
</tr>
</tbody>
</table>

-1.02 BLEU
Translation with Dependency Tree

- Actions predicted by using greedy search until "Shift" comes up
- Translation generated by using beam search with the beam size of 20

**Source:** 転移 温度 も 120 K 以上 は 実現 されて いない。
**Reference:** A transition temperature hasn’t been realized over 120K.

The dependency label error

The transition temperature has not been realized over 120 K. EOS (ROOT)
Conclusion

- Syntactic NMT model (NMT+RNNG) which learns to parse and translate
- Experimental results showed significant improvement over the baseline NMT
  - Our model generated a translation while parsing it
  - Stack in RNNG, where dependency trees are constructed, is a key component

Code: https://github.com/tempra28/nmtrnn