Character-based Decoding in Tree-to-Sequence Attention-based Neural Machine Translation

Akiko Eriguchi, Kazuma Hashimoto and Yoshimasa Tsuruoka (University of Tokyo); Team-ID: UT-AKY

1. Our System Submitted to WAT'16

Background

- Neural Machine Translation (NMT) models have achieved the state-of-the-art results in translation tasks.
- Syntax is shown to be useful for improving the NMT models, e.g. a tree-based encoder (Eriguchi et al., 2016) and feeding syntactic features to the encoder (Sennrich et al., 2016)

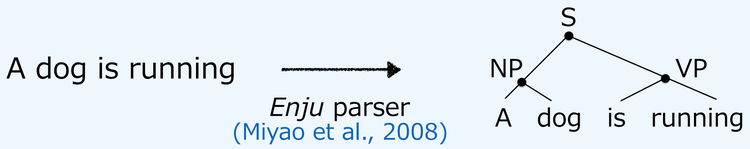

Figure 1: A parsed tree

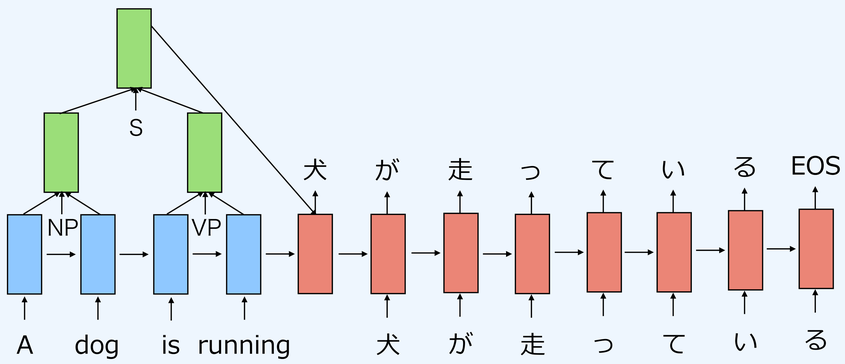

Figure 2: Overview of our system

What is the optimal units for NMT?

*Word-based Approach has a problem of **"UNK replacement"***

- Unknown words are required to be replaced with the appropriate translated words after translation

Sub-word-based Approach (Sennrich et al., 2016)

- Let the vocabulary size much smaller
- Unknown word

Character-based Approach (Chung et al., 2016)

- Explicit segmentation tool is unnecessary
- Almost all the words in a corpus can be covered

We investigate the effectiveness of the character-based decoding in the syntax-based NMT model (Eriguchi et al., 2016)

2. Experimental Setting and Results

	Word-based	Char-based
Hidden size:	512	512
Embedding size:	512	512 (src), 256 (tgt)
Calculation of softmax loss:	*BlackOut* sampling	the original softmax
mini-batch size:	128	128
Max of generated words:	100	300
Optimizer:	SGD	SGD
Training Time:	15 days	21 days
Beam size:	20	20

Table 1: Dataset

	Sentences	Parsed
Train	1,346,946	1,346,946
Development	1,790	1,789
Test	1,812	1,811

Table 2: Vocabulary size

	English	Japanese
word-based	87,796	65,680
char-based	87,796	3,004

Table 2: Translation Accuracy (JP: *KyTea*)

	BLEU	RIBES
Sequence-to-Sequence		
Word-based decoder	34.64 (67.5/43.9/30.3/21.5)	81.60
Tree-to-Sequence		
Word-based decoder	35.05 (68.1/44.6/31.0/22.1)	81.67
3 ensemble of the word-based	38.00 (70.4/47.7/34.1/25.0)	83.27
Char-based decoder	34.47 (65.3/42.3/29.2/20.7)	80.73
+ phrase label (64 dim.)	34.22 (64.4/41.8/28.8/20.4)	80.72
+ phrase label (128 dim.)	34.36 (65.9/42.9/29.7/21.1)	81.12
3 ensemble of the char-based	36.66 (67.7/45.3/32.1/23.4)	82.43
SMT		
Phrase-based baseline	29.80	69.19
Tree-to-string baseline	33.44	75.80

Translation Examples

Source sentence A:
The electric power generation was the 380 micro watt.
Ground truth A:
発電量は３８０マイクロワットであった。
Word-based:
発電は３８０UNKWであった。
Character-based:
発電は３８０マイクロワットであった。
(α = 0.78)

Source sentence B:
This paper describes development outline of low-loss forsterite porcelain.
Ground truth B:
低損失フォルステライト磁器の開発概要などを述べた。
Word-based:
ここでは，UNKUNKの開発概要を述べた。
Character-based:
低損失フォルステライト磁器の開発概要を述べた。

3. Discussion and Conclusion

- Character-based decoder can translate the word "380" into "3 8 0" without copy mechanism (Ling et al., 2016), because their data are included in the training dataset
- UNK replacement is usually required by word-based decoder but not by character-based decoder

Table 3: Average time of decoding

	Time (msec/ sentence)
Word-based	363.7
Char-based	8.8

Character-based decoder does not outperform the word-based decoder but exhibits two promising properties:

1) It takes much less time to compute the softmax layer 2) It can translate any word in a sentence