Tree-to-Sequence Attentional Neural Machine Translation

Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka (The University of Tokyo)

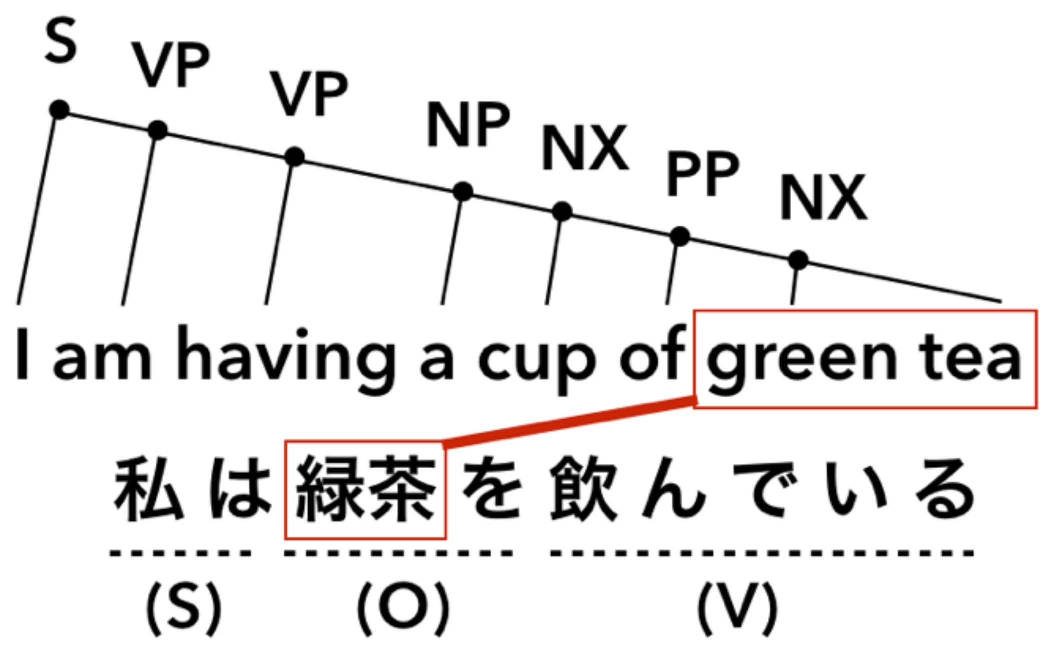
Introduction

✦ Neural Machine Translation (NMT) models have recently achieved state-of-the-art results on translation tasks (Luong et al., 2015)

“A B C <eos>” → Encoder → Decoder → “X Y Z <eos>”

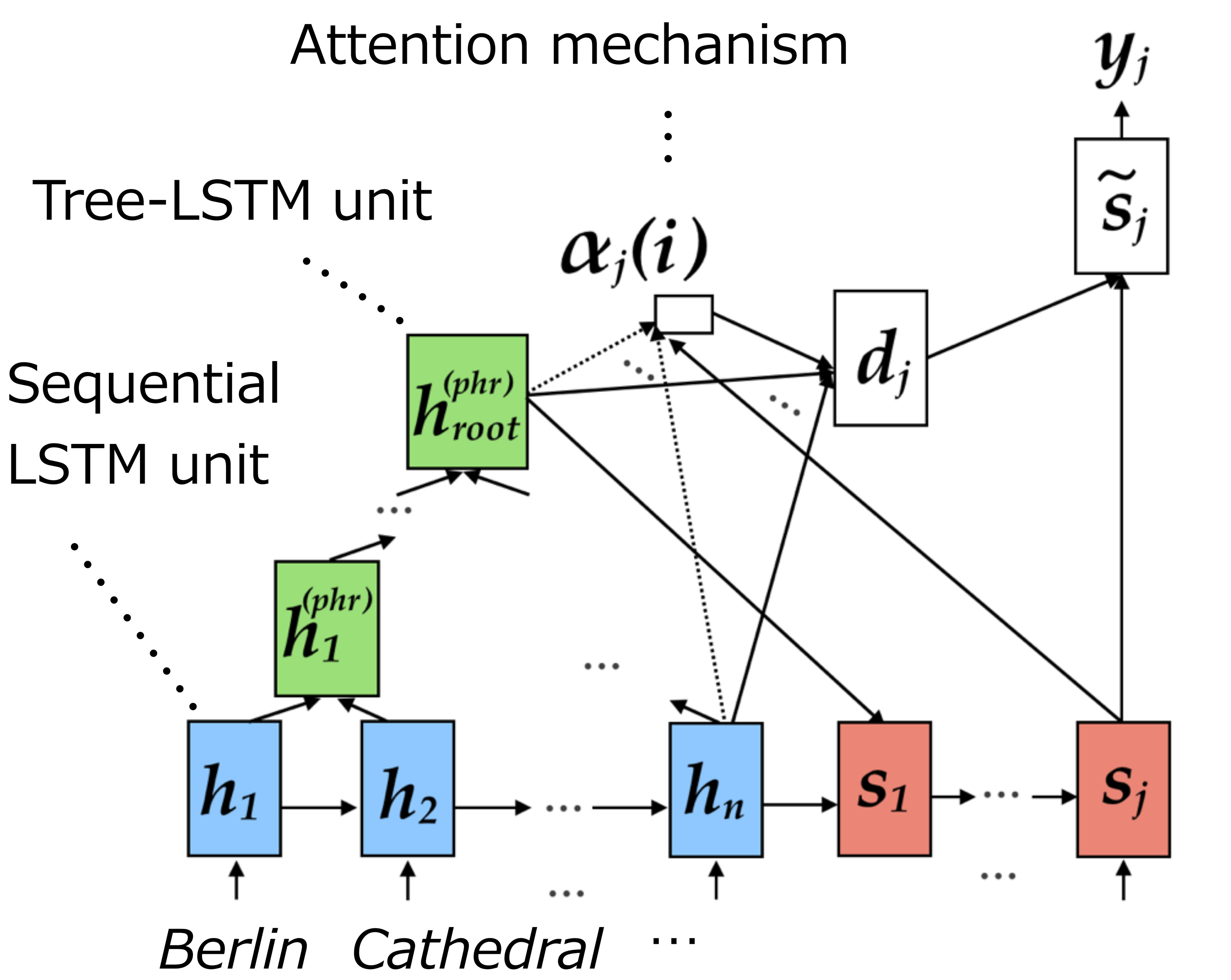
✦ In statistical machine translation, it is well known that incorporating syntax information improves translation accuracy (Liu et al., 2006)

✦ We proposed a novel tree-based NMT model and confirmed that our model achieved state-of-the-art result



- phrase structure
- word-to-phrase alignment

Proposed Model: *Tree-to-Sequence Attentional NMT Model*



Experimental results on a small EN-JP dataset of WAT'15

(i) Tree-based Encoder

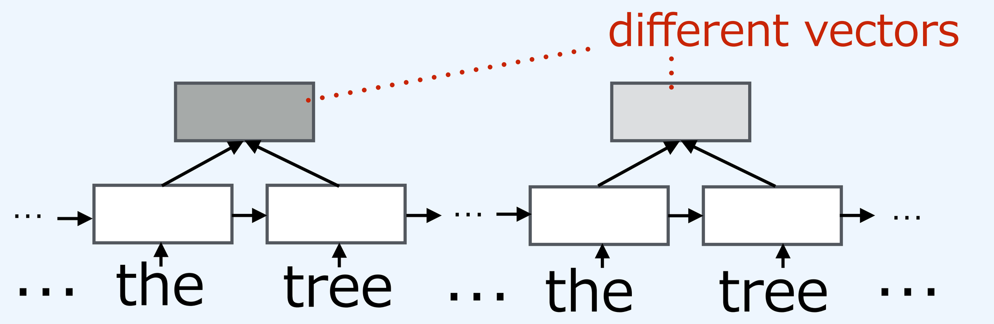
- **Phrase vectors** calculated by Tree-LSTMs (Tai et al., 2015)

	BLEU
Our model	**20.5**
ANMT model (Luong et al., 2015)	19.4
+ input reverse	17.5

Tree-based encoder outperforms sequence-based encoder

- **Leaf vectors** replaced by sequential LSTM units

	BLEU
with sequential LSTMs	**20.0**
w/o sequential LSTMs	19.5



Our model can distinguish the same phrases in a sentence

(ii) Attention to the Source Tree

- softly aligning a target word with source words / phrases

(iii) BlackOut (Ji et al., 2015)

- reduces computational cost of *softmax* on **CPU**s

	BLEU	Time/ Epoch
BlackOut	20.5	**70 min.**
Original softmax	21.8	180 min.

Experimental Results

	RIBES*	BLEU
Our model (d = 512)	81.46	34.36
Our model (d = 768)	81.89	34.78
Our model (d = 1024)	81.58	34.87
Ensemble of the above 3 models	**82.45**	36.95
ANMT model (Luong et al., 2015)	80.64	32.88
ANMT model (Ensemble and unk replacement) on GPU (Zhu, 2015)	79.70	34.19
ANMT model (+ character-based decoding) on GPU (Lee et al., 2015)	81.15	35.75
Phrase-based baseline	69.19	29.80
Hierarchical phrase-based baseline	74.70	32.56
Tree-to-string baseline	75.80	33.44
Tree-to-string (Neubig and Duh, 2014)	79.65	36.58
+ ANMT Rerank (Neubig et al., 2015)	81.38	38.17

*Evaluation metrics: The RIBES score (Isozaki et al., 2013) has higher correlation with human judgement in EN-JP translation than BLEU (Papienini et al., 2013).

✦ **Dataset: WAT'15 (EN-JP Task; 1.3M)**

- Vocabulary: (EN, JP) = (87K, 65K)
- Parser: *Enju* (Miyao et al., 2008)

✦ **Model Parameters**

- Hidden size: d ∈ {512, 768, 1024}
- Embed size: 512
- BlackOut: negative samples = 2500

Code is available at https://github.com/tempra28/tree2seq

Translation Example

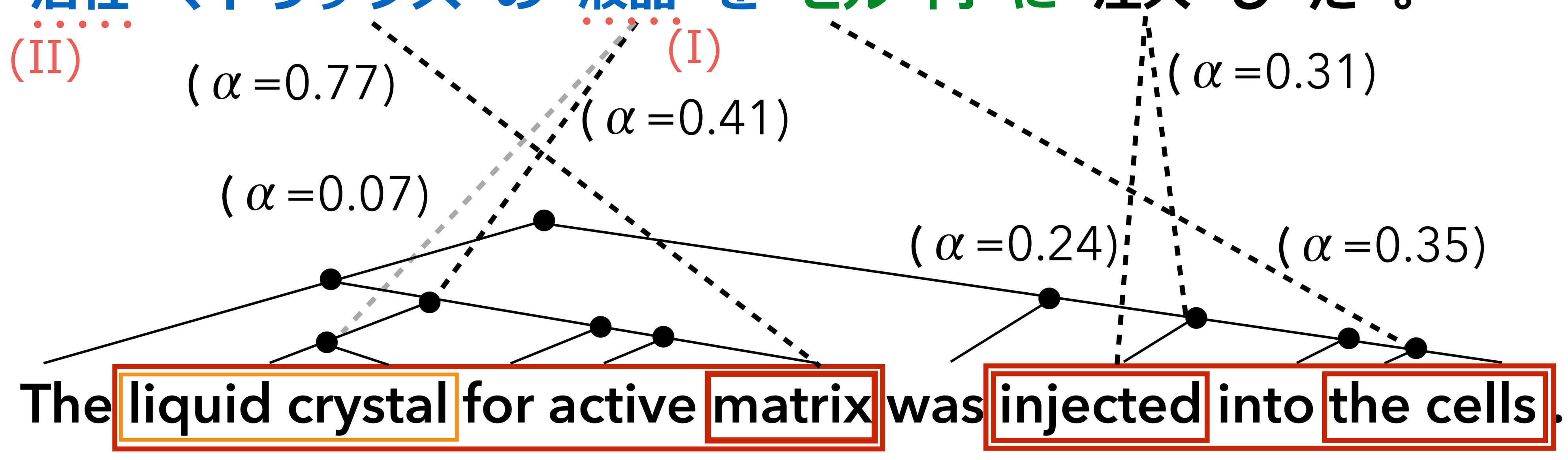
[Reference]

セル に は アクティブ マトリックス 用 液晶 を 注入 し た 。

(II)

[Translation]

活性 マトリックス の 液晶 を セル 内 に 注入 し た 。



(I) **Effects of Attention mechanism:** “マトリックス” (“matrix” in English) has the highest attention score with the word “matrix”. “液晶” (“liquid crystal” in English) has a higher attention score with the phrase “liquid crystal for active matrix”.

(II) **Synonymous translation**: “アクティブ” and “活性”, both of which means “active” in English.