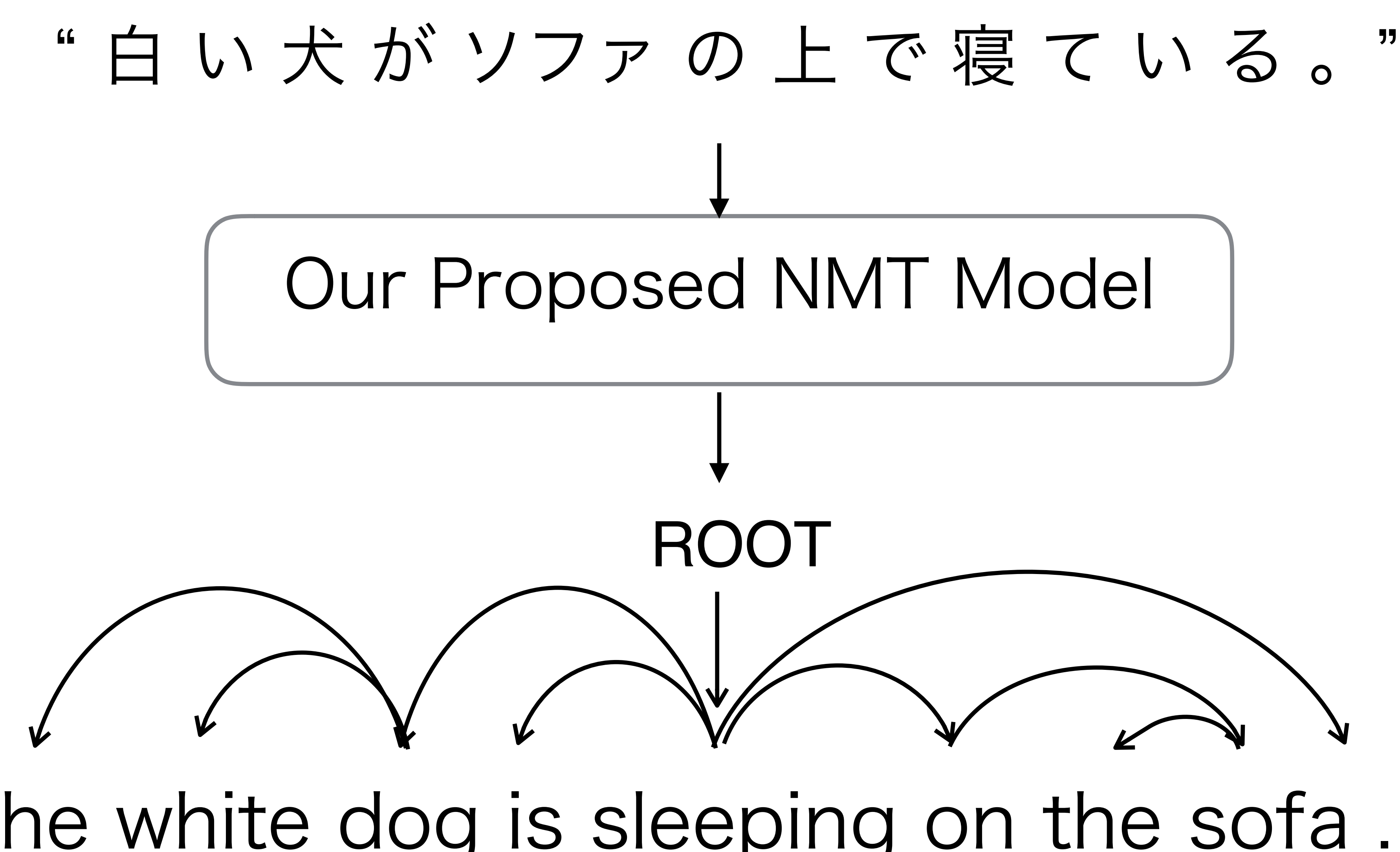
Learning to Parse and Translate Improves Neural Machine Translation

Akiko Eriguchi[1], Yoshimasa Tsuruoka[1], and Kyunghyun Cho[2]
[1] The University of Tokyo
[2] New York University

東京大学 THE UNIVERSITY OF TOKYO NYU

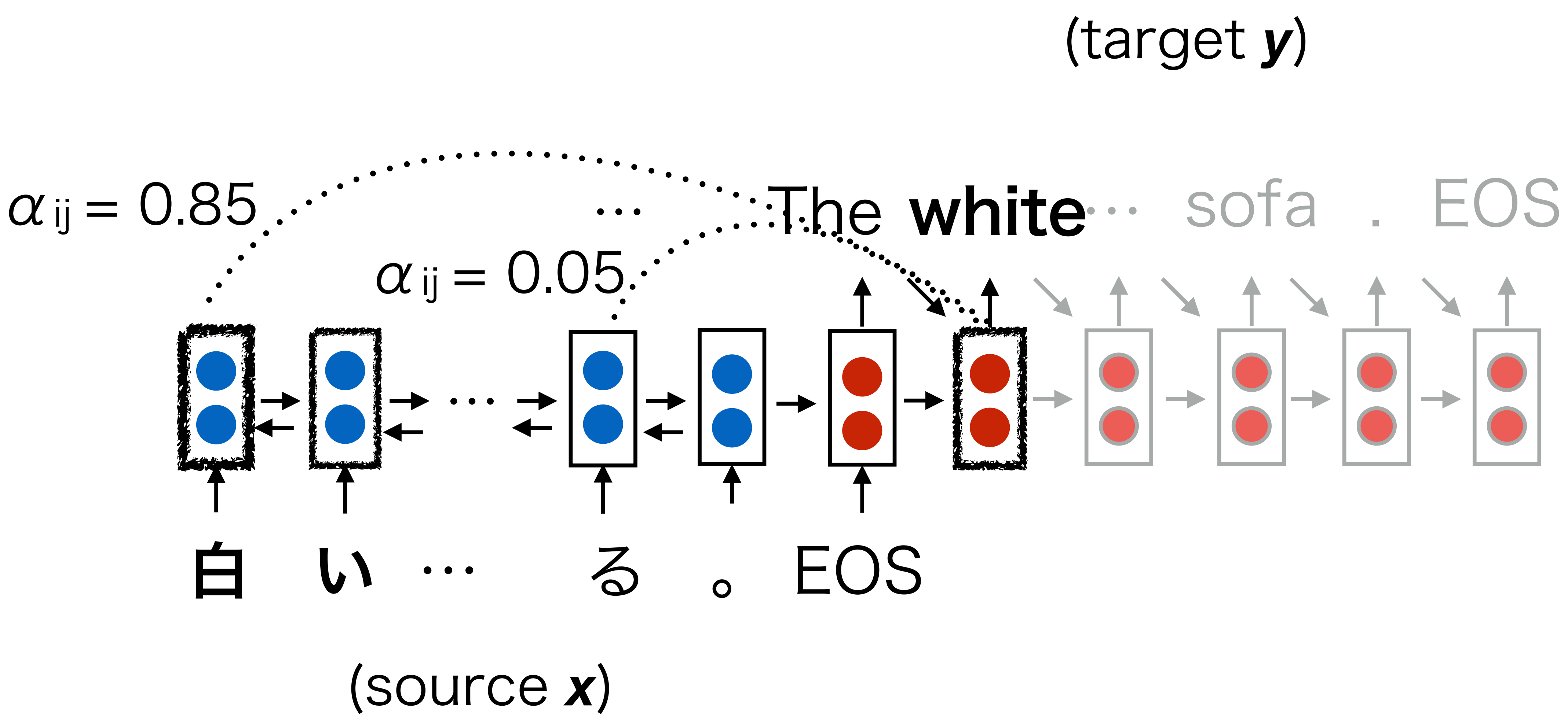
Overview

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



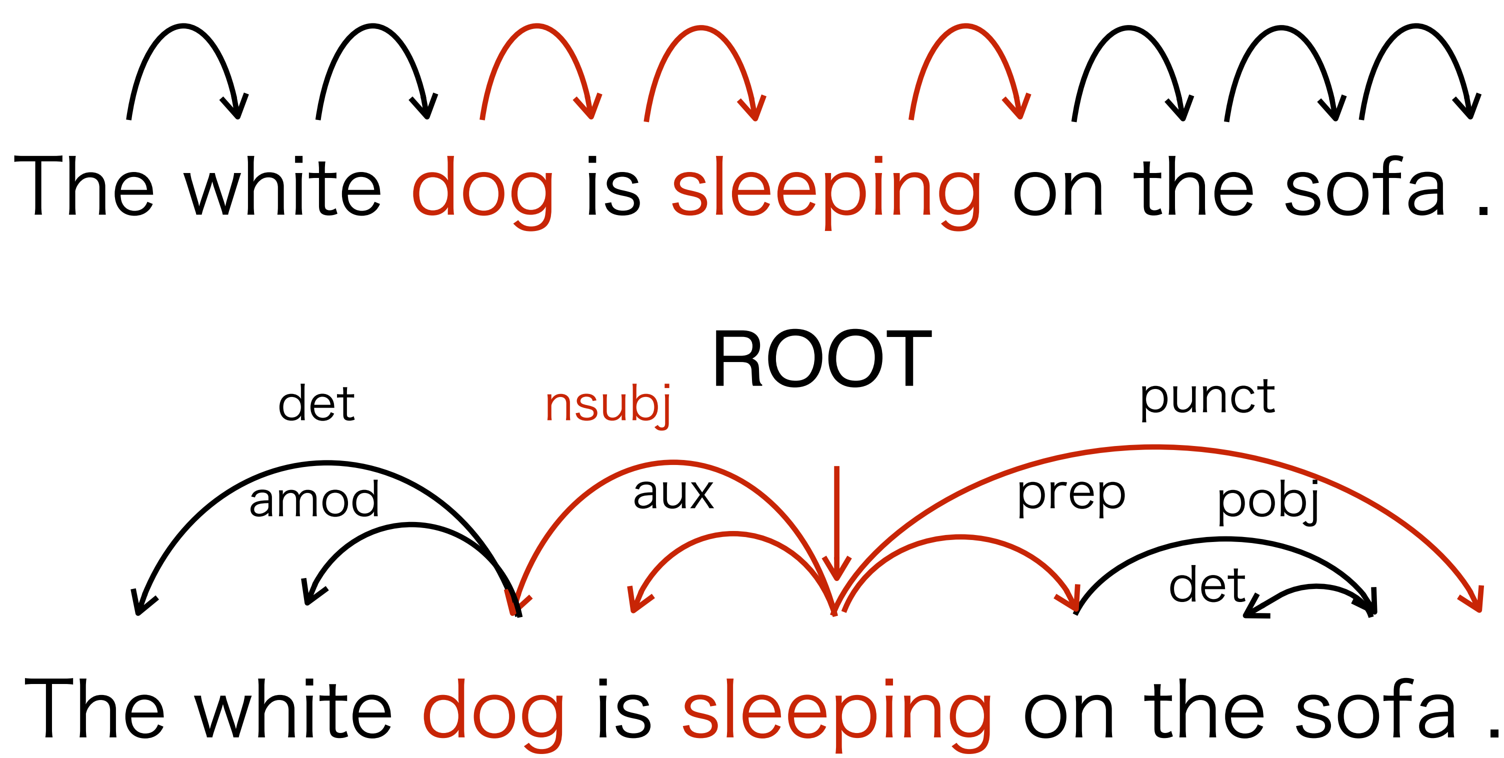
Neural Machine Translation

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



Motivation

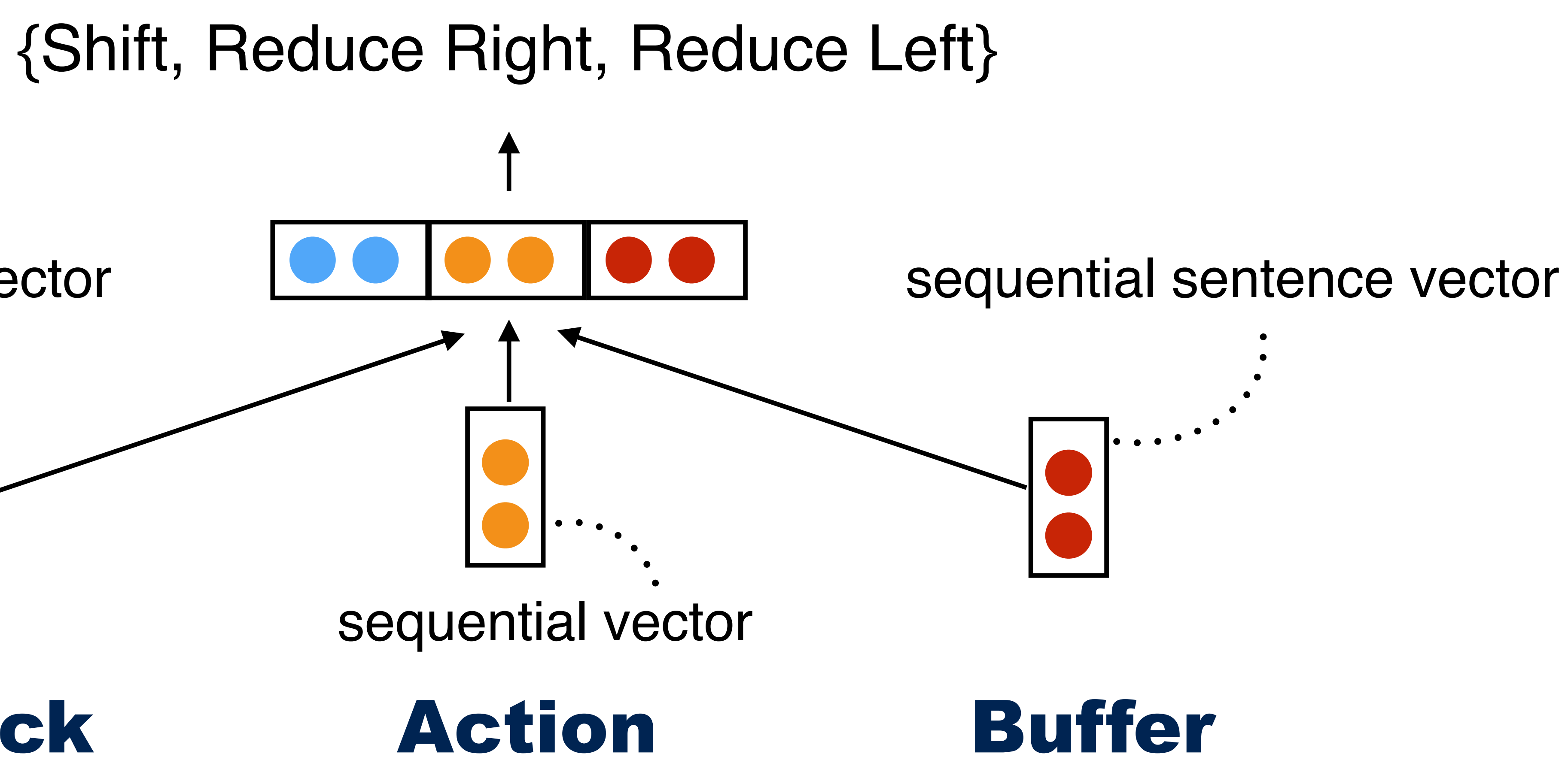
- RNN-based NMT model captures a sequence of words and does NOT utilize syntactic information explicitly
- To decode a translation with its parsed tree is expected to be useful (e.g. in selecting a grammatical sentence)



Recurrent Neural Network Grammars (RNNG)

(Dyer et al., 2016)

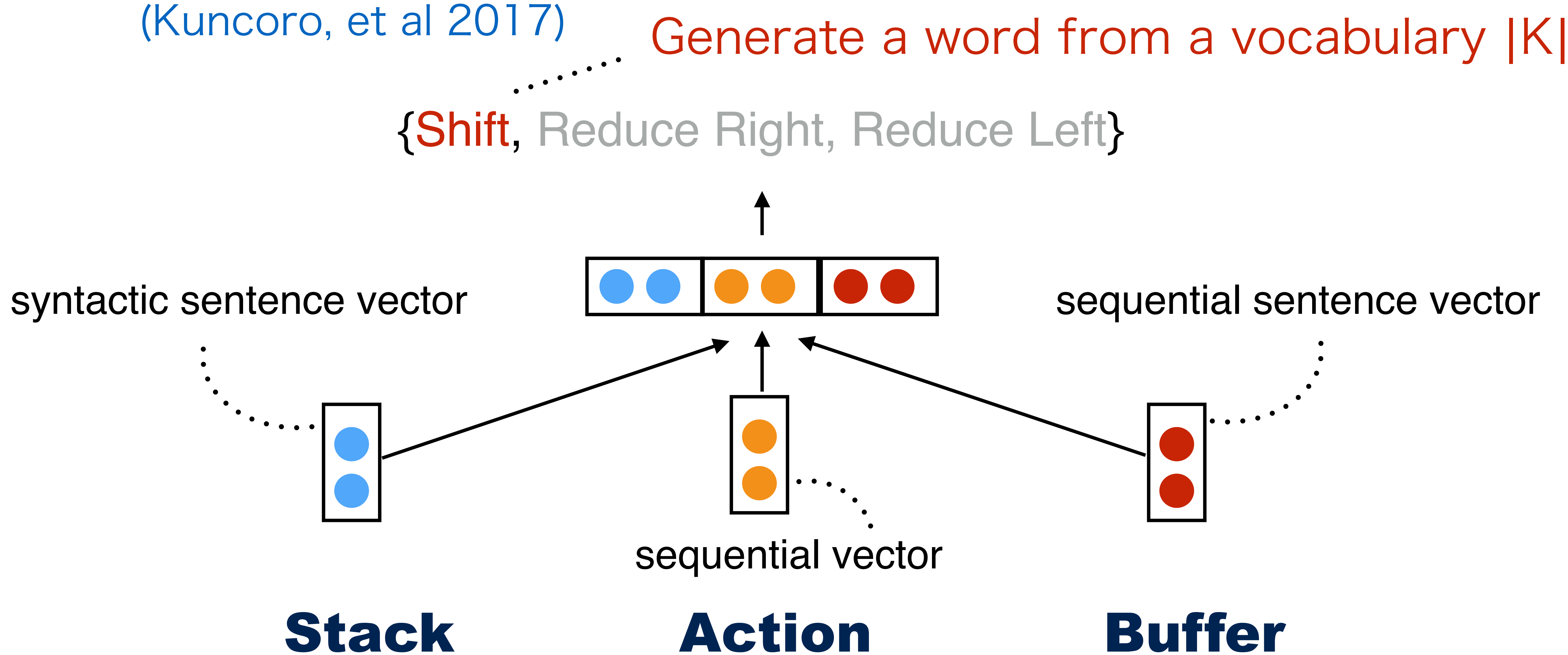
- Joint model of shift-reduce parsing and language model
  - Selects an action out of three types of actions from the composed vectors of Stack, Action, and Buffer
  - Achieved the state-of-the-art performance in both tasks (Kuncoro et al., 2017)



Recurrent Neural Network Grammars (RNNG)

(Dyer et al., 2016)

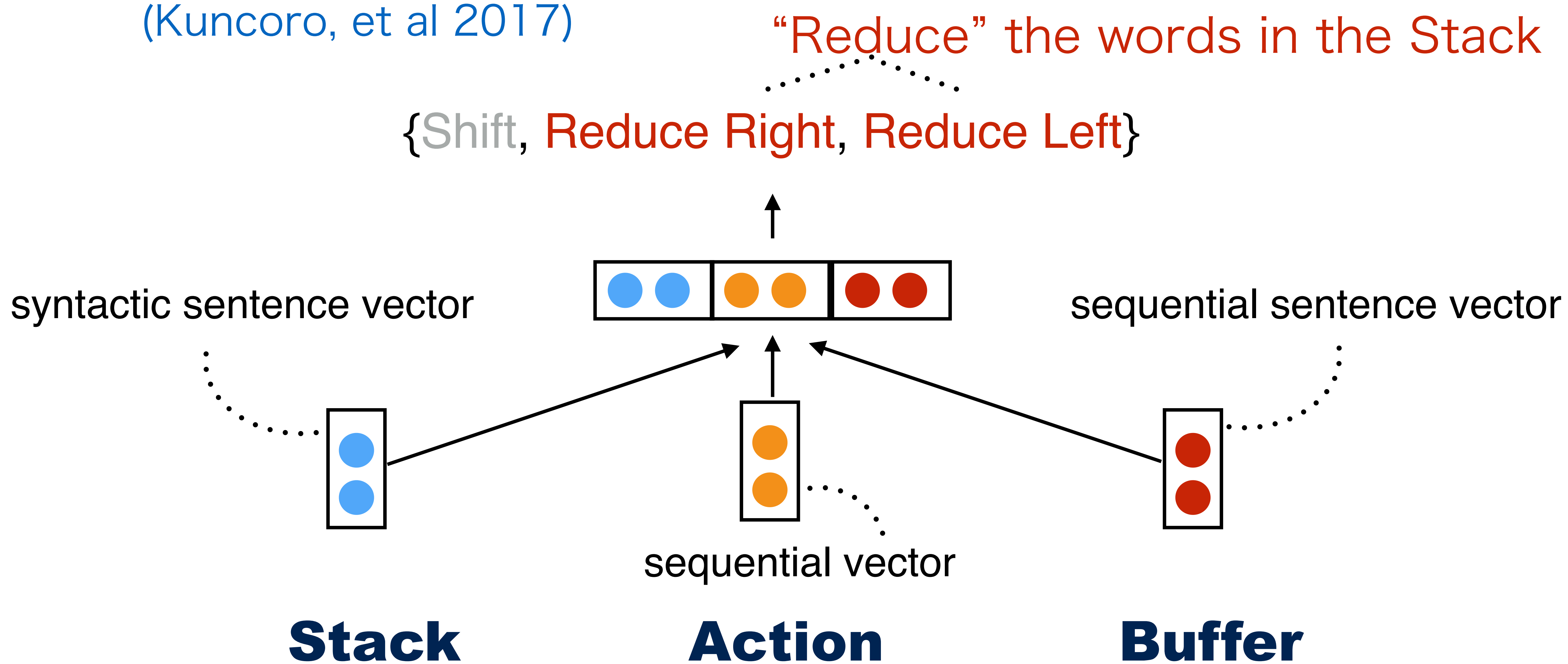
- Joint model of shift-reduce parsing and language model
  - Selects an action out of three types of actions from the composed vectors of Stack, Action, and Buffer
  - achieved the state-of-the-art performance in both tasks (Kuncoro, et al 2017)



Recurrent Neural Network Grammars (RNNG)

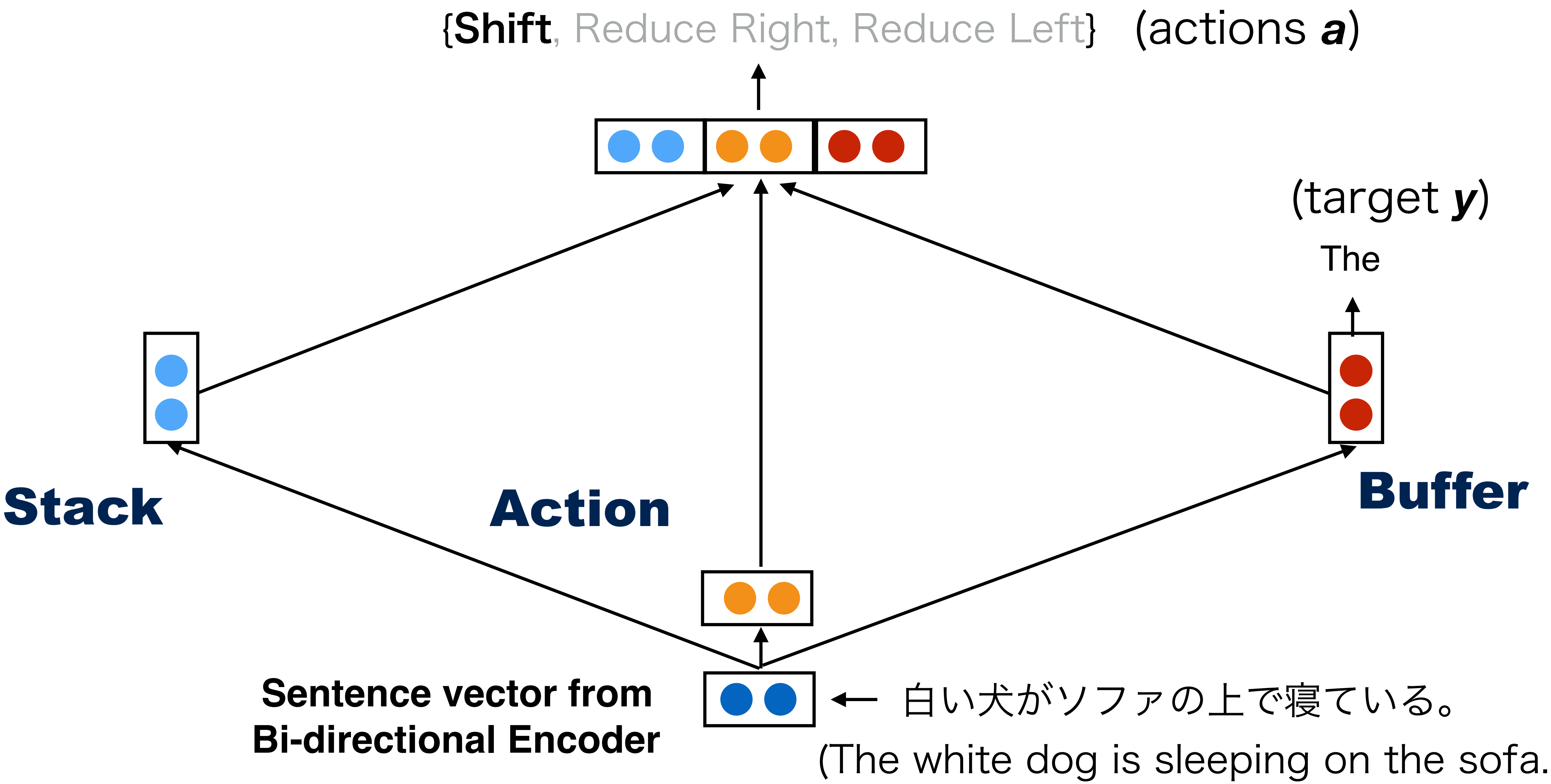
(Dyer et al., 2016)

- 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)



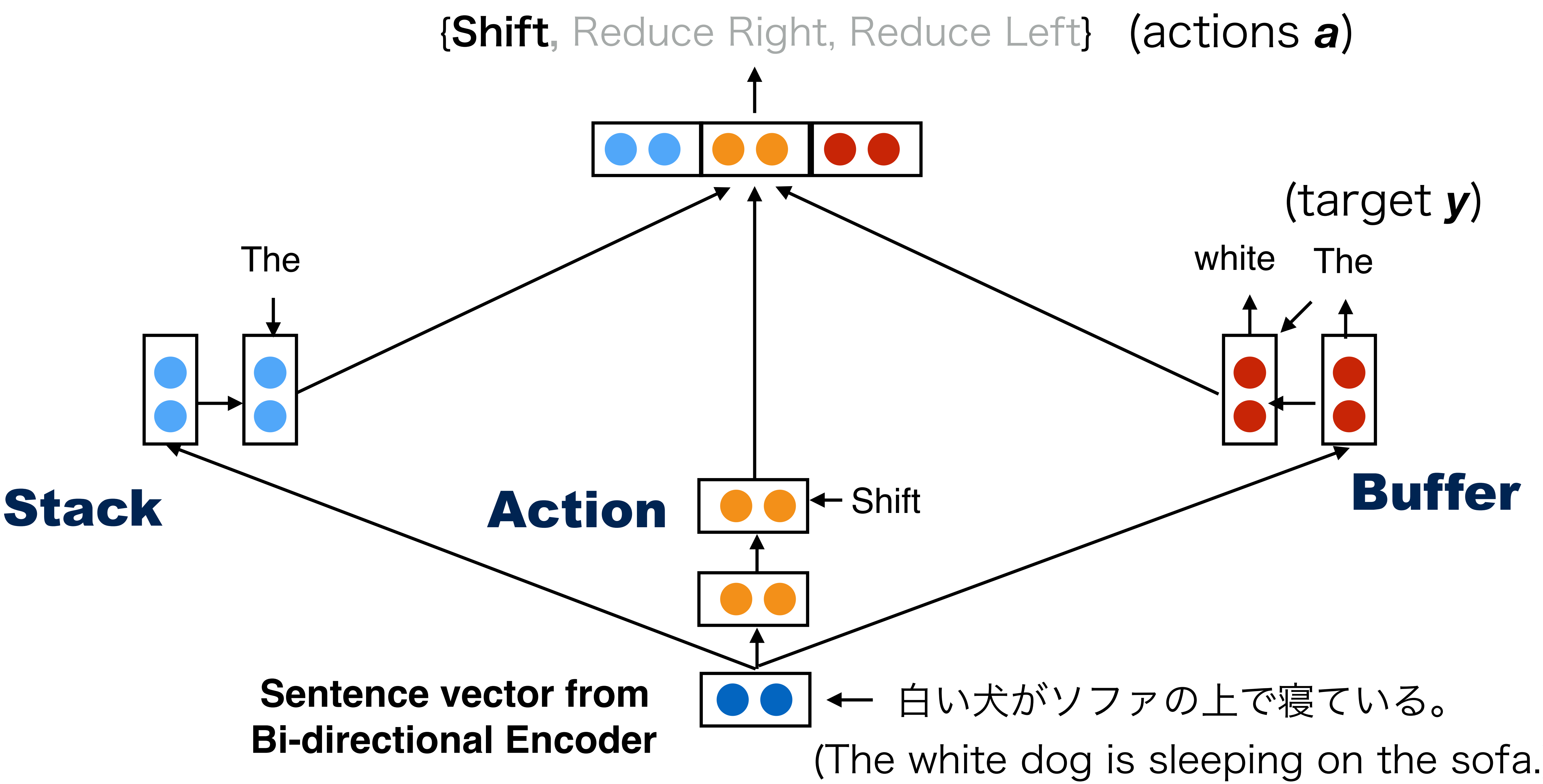
Proposed Model: NMT + RNNG

- Objective function: $J(\Theta) = \Sigma \log p(\boldsymbol{y}, \boldsymbol{a} | \boldsymbol{x})$ (***y***'s parsed actions ***a***)



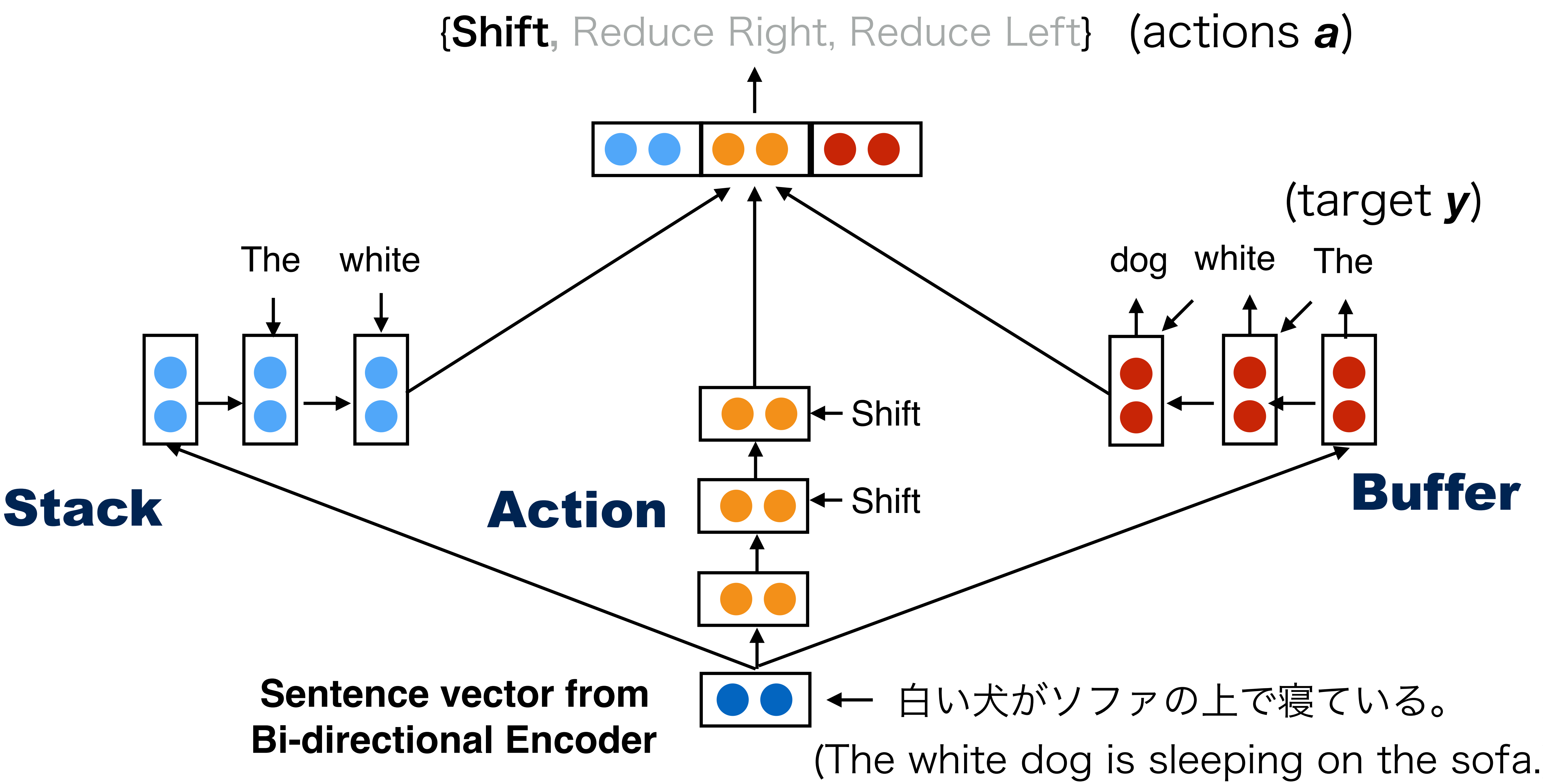
Proposed Model: NMT + RNNG

- Objective function: $J(\Theta) = \Sigma \log p(\mathbf{y}, \mathbf{a} \mid \mathbf{x})$ (***y***'s parsed actions ***a***)



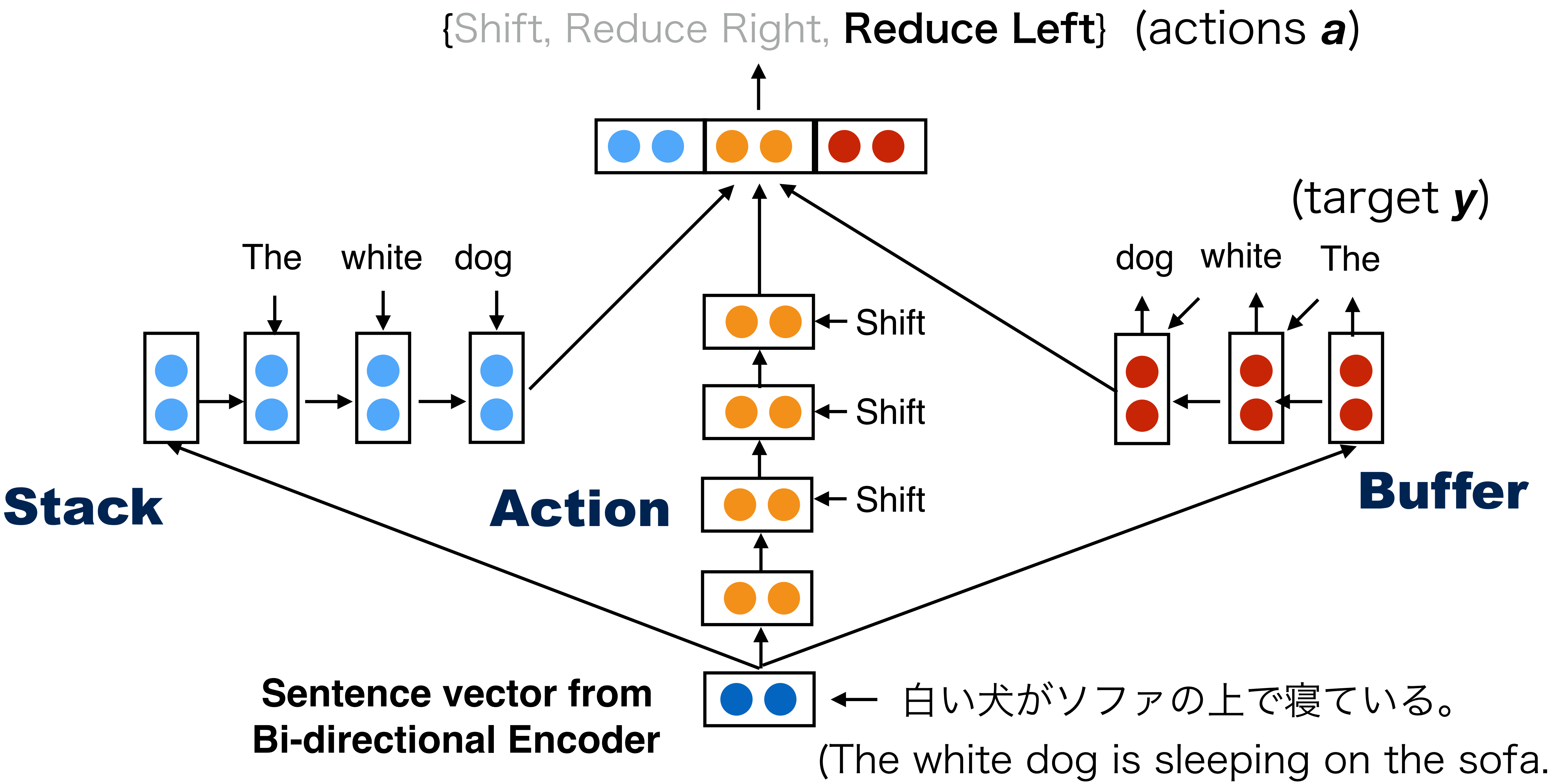
Proposed Model: NMT + RNNG

- Objective function: $J(\Theta) = \Sigma \log p(\mathbf{y}, \mathbf{a} | \mathbf{x})$ (***y***'s parsed actions ***a***)



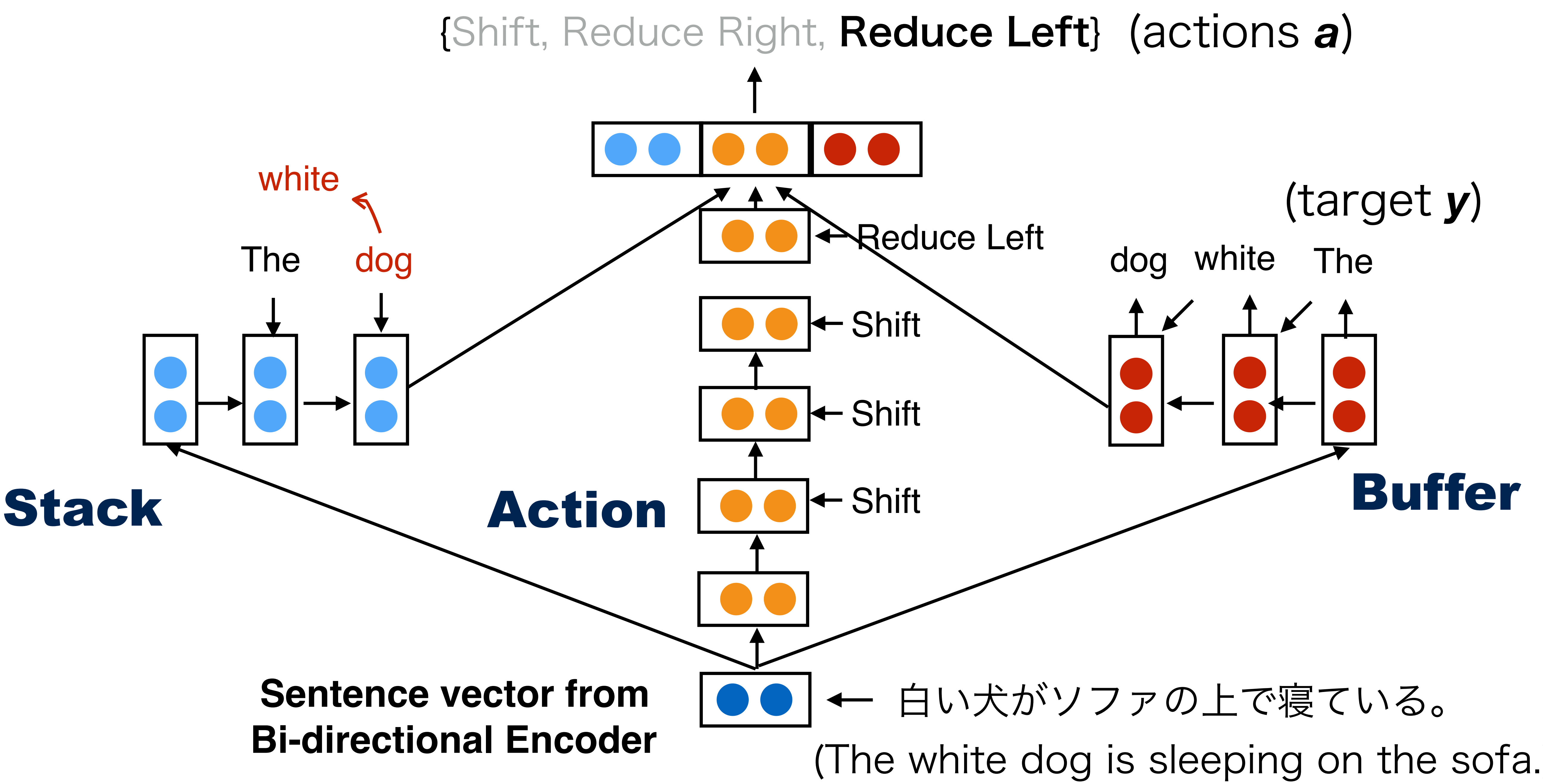
Proposed Model: NMT + RNNG

- Objective function: $J(\Theta) = \Sigma \log p(\mathbf{y}, \mathbf{a} | \mathbf{x})$ (***y***'s parsed actions ***a***)



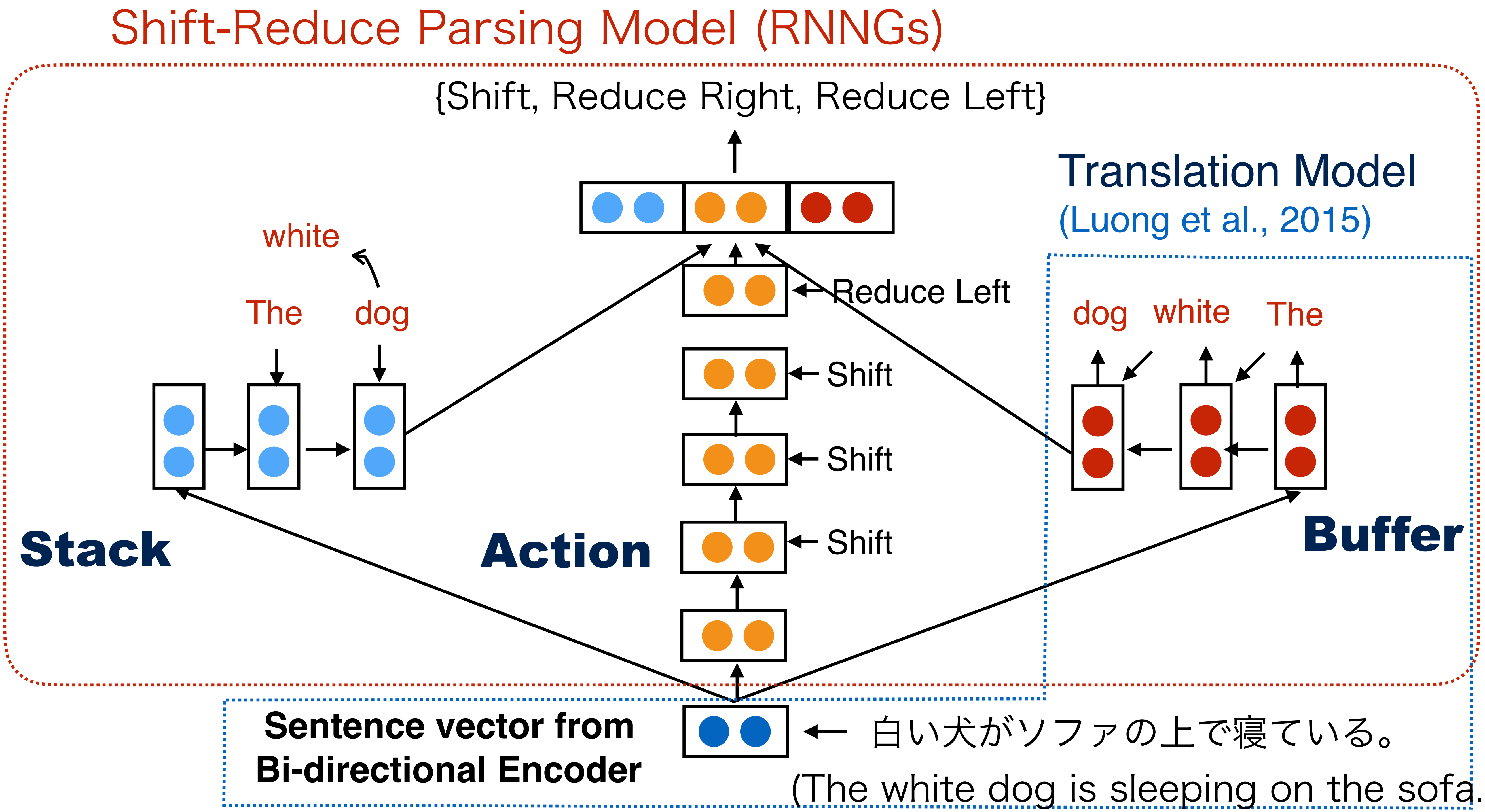
Proposed Model: NMT + RNNG

- Objective function: $J(\Theta) = \Sigma \log p(\boldsymbol{y}, \boldsymbol{a} | \boldsymbol{x})$ (***y***'s parsed actions ***a***)



Proposed Model: NMT + RNNG

- Objective function: $J(\Theta) = \Sigma \log p(\boldsymbol{y}, \boldsymbol{a} | \boldsymbol{x})$ (***y***'s parsed actions ***a***)



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

	JP-EN	RU-EN	CS-EN	DE-EN
NMT	17.88	12.03	11.22	16.61
NMT + RNNG	※**18.84**	※**12.46**	※**12.06**	16.41

RIBES

	JP-EN	RU-EN	CS-EN	DE-EN
NMT	71.27	69.56	69.59	73.75
NMT + RNNG	※**72.25**	※**71.04**	※**70.39**	※**75.03**

※ significant difference by bootsrap 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)

BLEU

	JP-EN (Dev.)
NMT + RNNG	**18.60**
w/o Buffer	18.02
w/o Action	17.94
w/o Stack	**17.58**
NMT	17.75

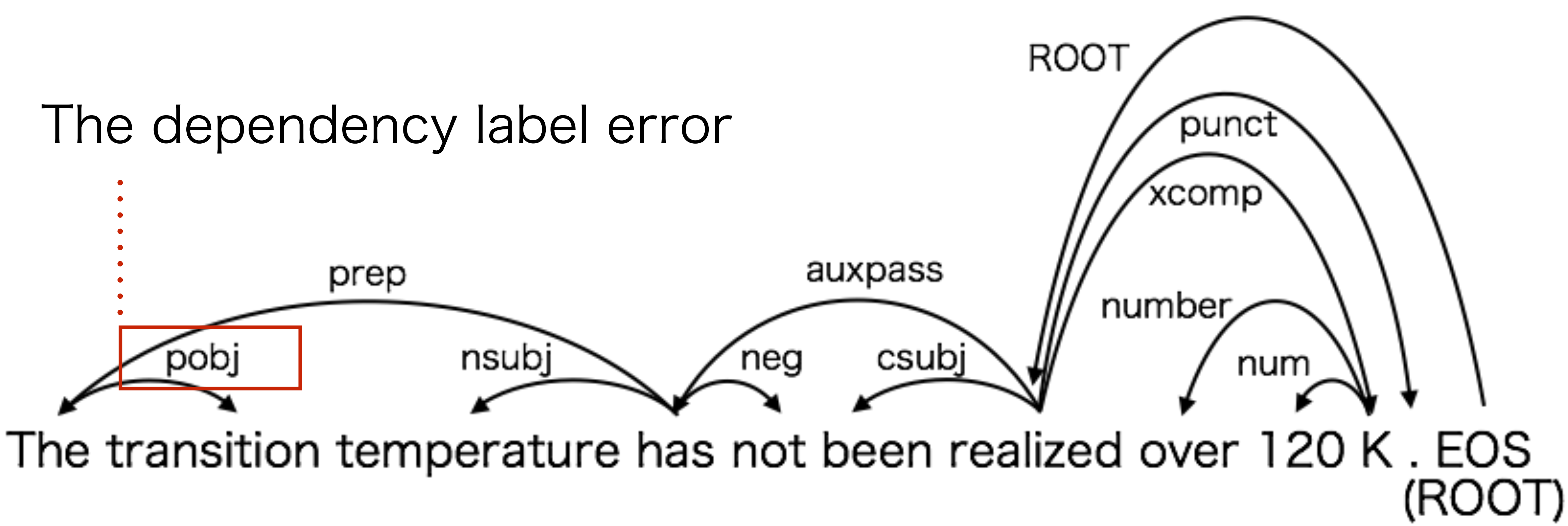
-1.02 BLEU

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 １ ２ ０ Ｋ 以上 は 実現 さ れ て い な い 。

Reference: A transition temperature hasn't been realized over 120K .



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/nmtrnng