Simultaneous Neural Machine Translation with Prefix Alignment

Yasumasa Kano, Katsuhito Sudoh, Satoshi Nakamura

Nara Institute of Science and Technology (NAIST), Japan

Simultaneous translation

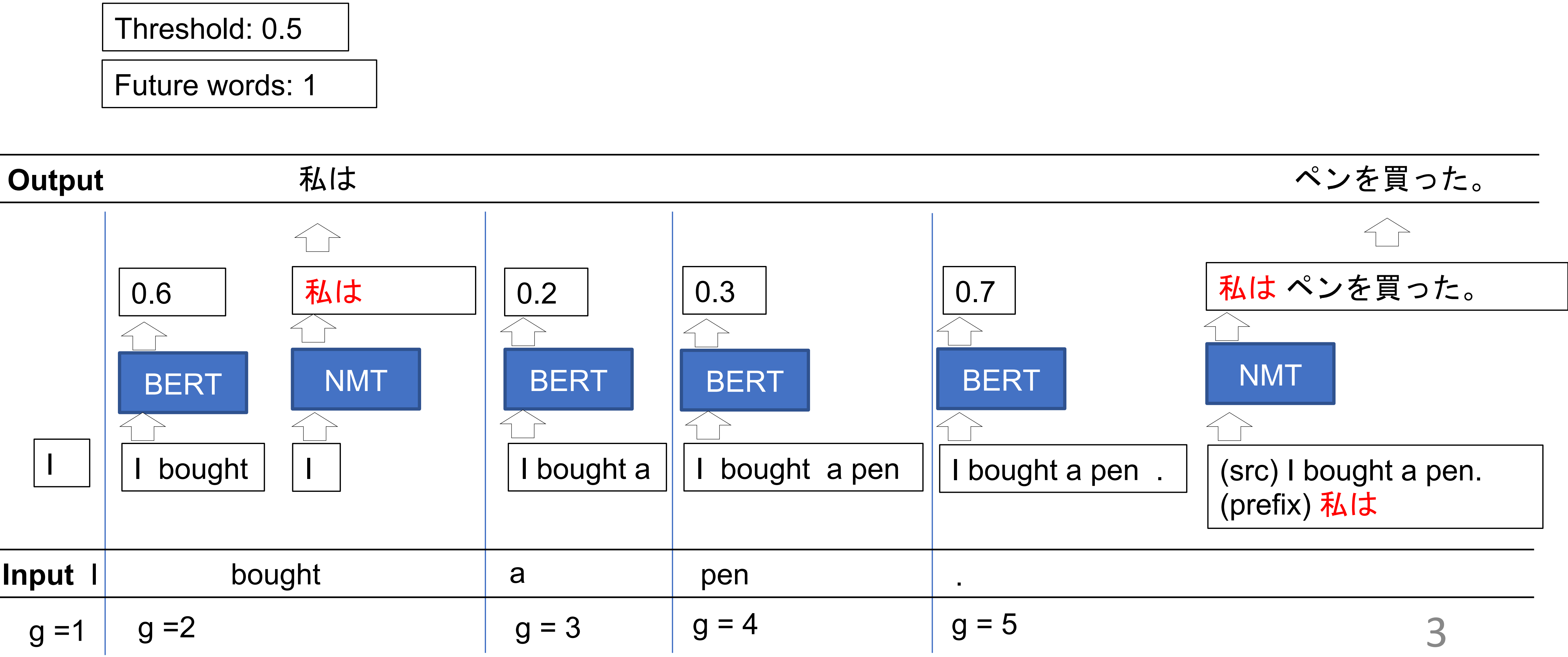
- Consecutive translation

Input:	I	am	a	student	.					
Output:						私	は	学生	です	。

- Simultaneous translation

Input:	I	am	a	student	.
Output:	私	は	学生	です	。

Related work: Meaningful Unit [Zhang+, 2020]



Problem

- The previous work: NMT model trained with full sentences.

Input	Output
I	私です。
I bought a pen	私です。ペンを買った。

- Proposed method: NMT model fine-tuned with bilingual prefix pairs

Input	Output
I	私
I bought a pen	私はペンを買った。

How to extract bilingual prefix pairs

1. Translate the source prefix with pre-trained NMT model

Source Prefix	**Source prefix Translation**	**Full-sentence translation**	**Extracted Target Prefix**
I	僕は。	僕はペン買った。	僕は
I bought	僕は買った。	僕はペンを買った。	
I bought a	僕は買った。	僕はペンを買った。	
I bought a pen	僕はペンを買った	僕はペンを買った。	僕はペンを買った
I bought a pen .	僕はペンを買った。	僕はペンを買った。	僕はペンを買った。

How to extract bilingual prefix pairs

2. Find reference prefixes corresponding to the prefix translation pairs

2.1. Extracted Pairs
(source, translation prefix)
(I, 僕は) → 僕は
(I bought a pen, 僕はペンを買った)

2.2.Calculate BERT score with reference prefixes

0.6	私
0.8	私は
0.3	私はペン
0.2	私はペンを
0.1	私はペンを買った

2.3. Prefix pairs with reference
(source, reference prefix)
(I, 私は)
(I bought a pen, 私はペンを買った)

Use the extracted prefix pairs

- Fine-tune the NMT model

  Data
  (I, 私は)
  (I bought a pen, 私はペンを買った)

- Train Boundary Predictor (binary classifier)

  Data
  (I, 1)
  (I bought, 0)
  (I bought a, 0)
  (I bought a pen, 1)

Experiment of simultaneous translation

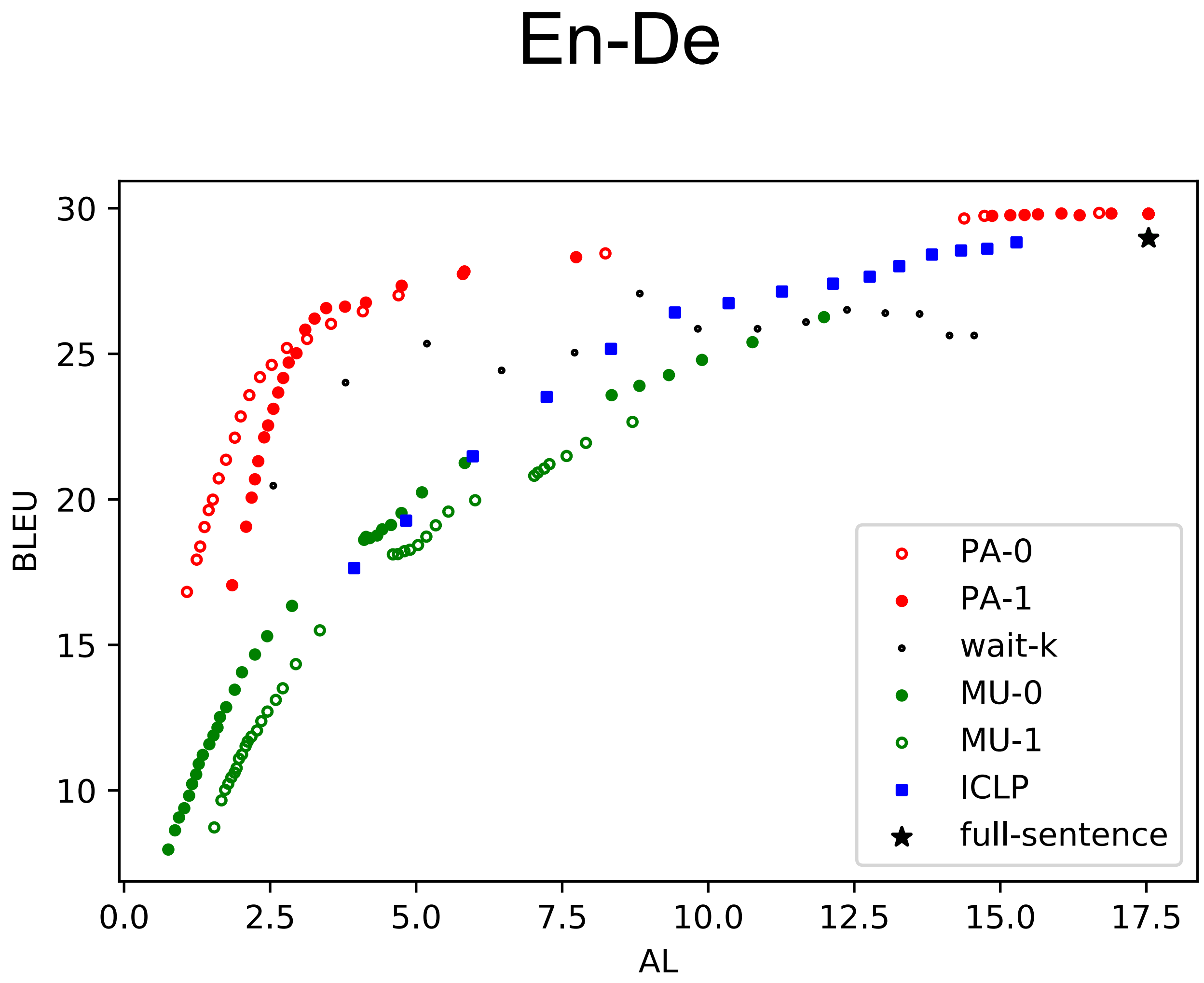
]

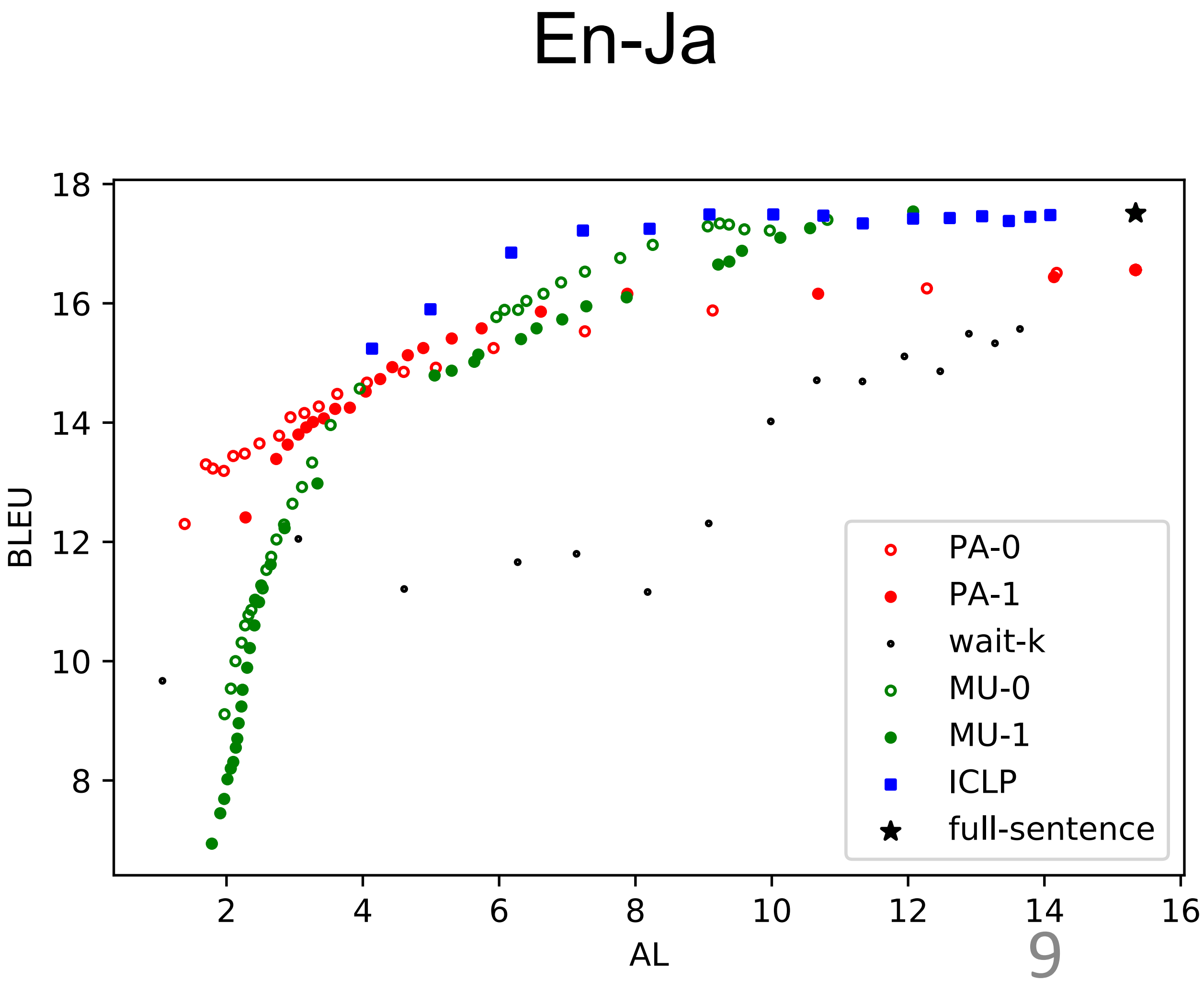
- Data

	En-De	En-Ja
Pre-train	4.5M (WMT2014)	20 M (WMT2020)
Fine-tune	206 K (IWSLT2021)	200 K (IWSLT2021)
Dev	5.6 K (IWSLT 2017 dev-test)	5.3 K (IWSLT 2017 dev-test)
Test	1.0 K (IWSLT2015 test)	1.5 K (IWSLT2021 dev)

- Subwords: Joint vocabulary size 16k (BPE)
- NMT Model: Transformer [Vaswani+, 2017]
- Boundary Predictor: BERT [Devlin+,2019]
- Evaluation metrics
    - Quality: BLEU
    - Latency: AL (Average Lagging) [Ma+ , 2019]

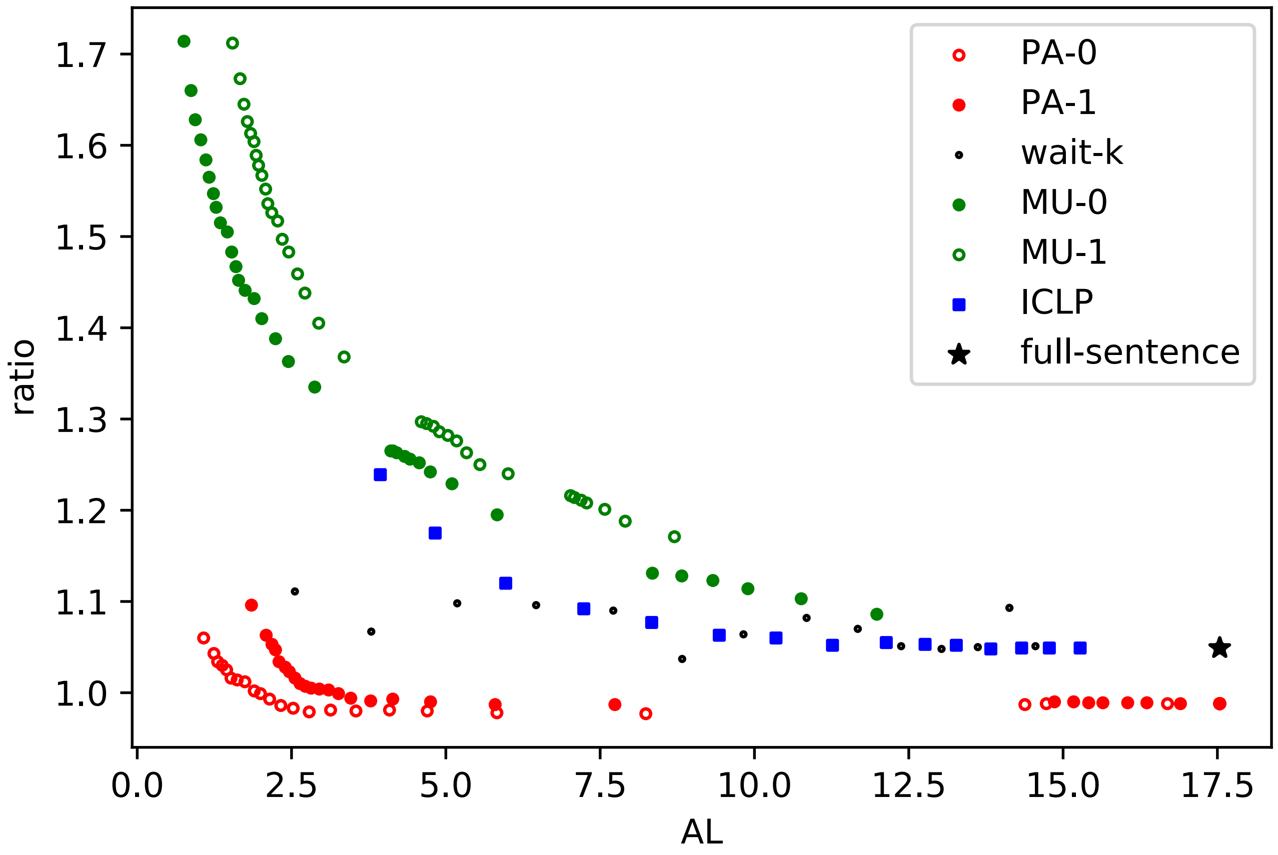
Result (BLEU)


En-De
BLEU
AL
30
25
20
15
10
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
PA-0
PA-1
wait-k
MU-0
MU-1
ICLP
full-sentence

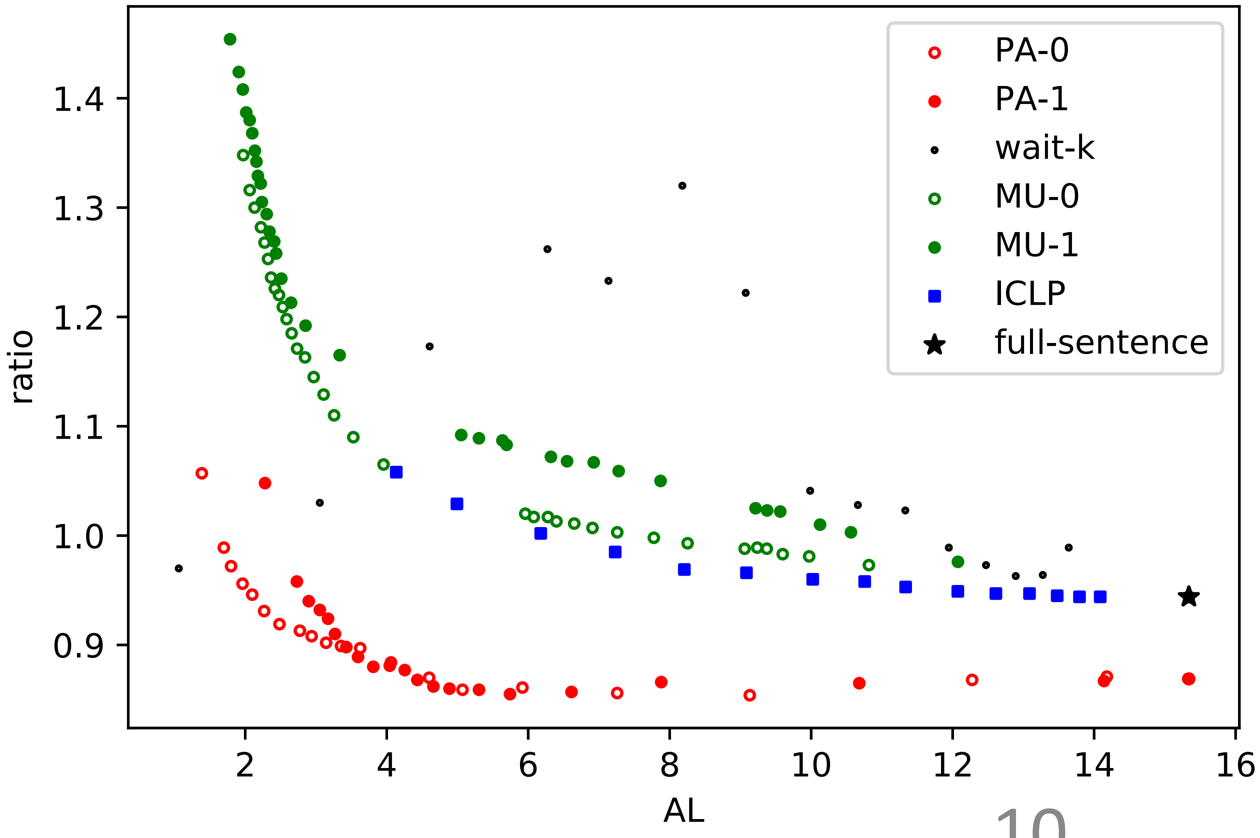

En-Ja
BLEU
AL
18
16
14
12
10
8
2
4
6
8
10
12
14
16
PA-0
PA-1
wait-k
MU-0
MU-1
ICLP
full-sentence

Result (Length Ratio)

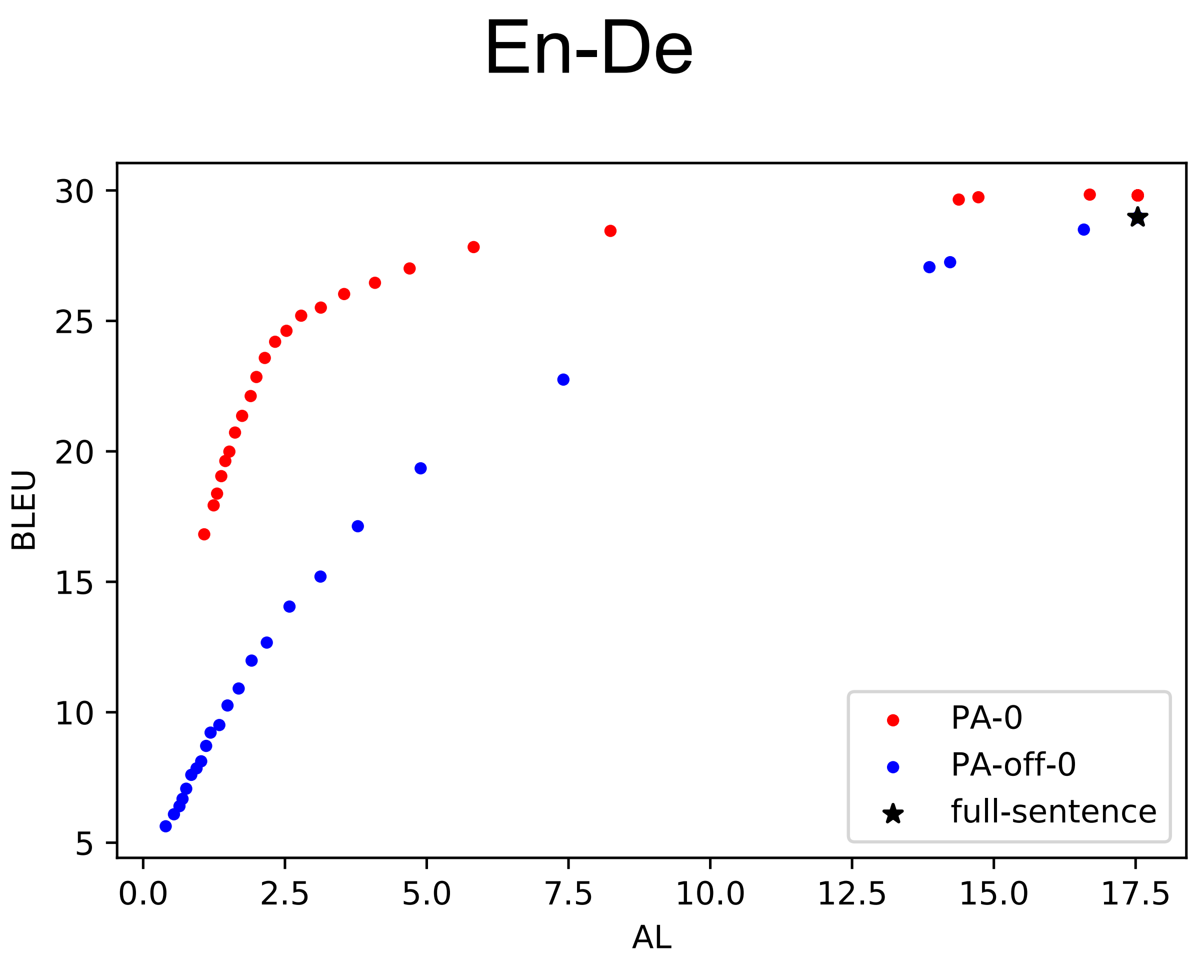
En-De

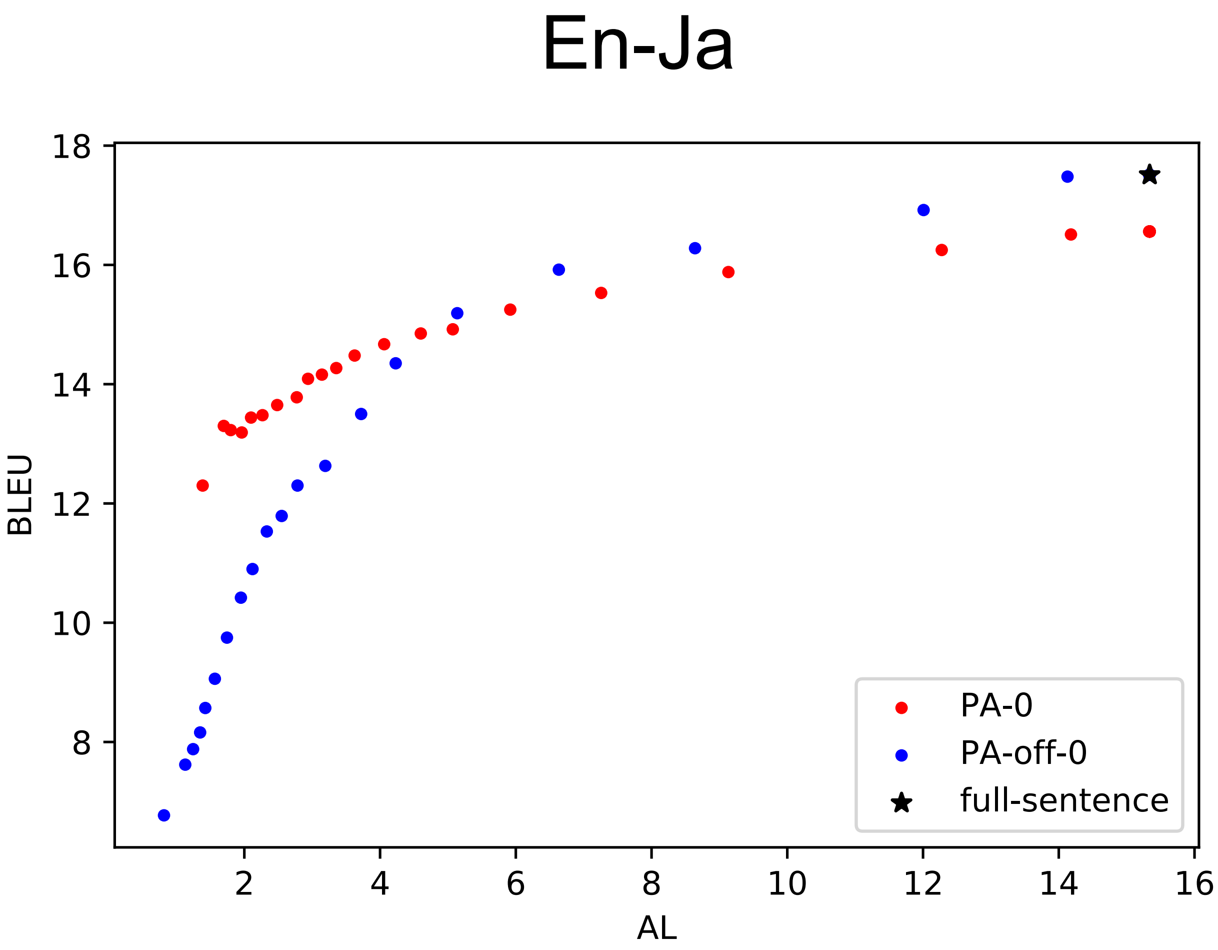


En-Ja



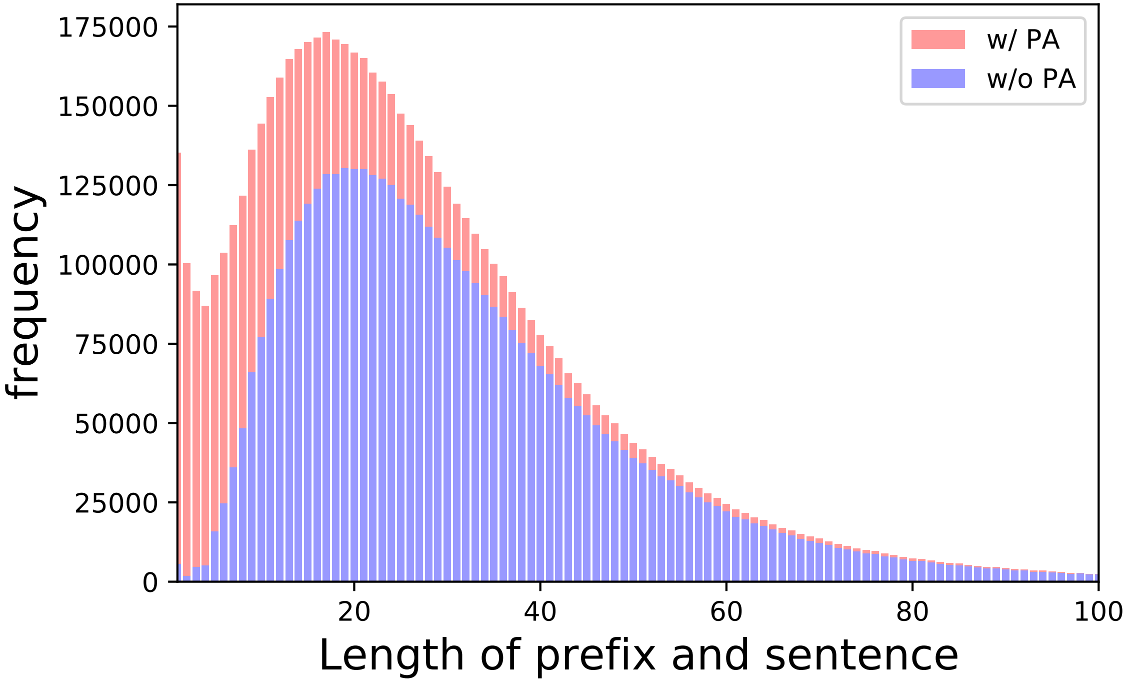
Effect of Fine-tuning (BLEU)



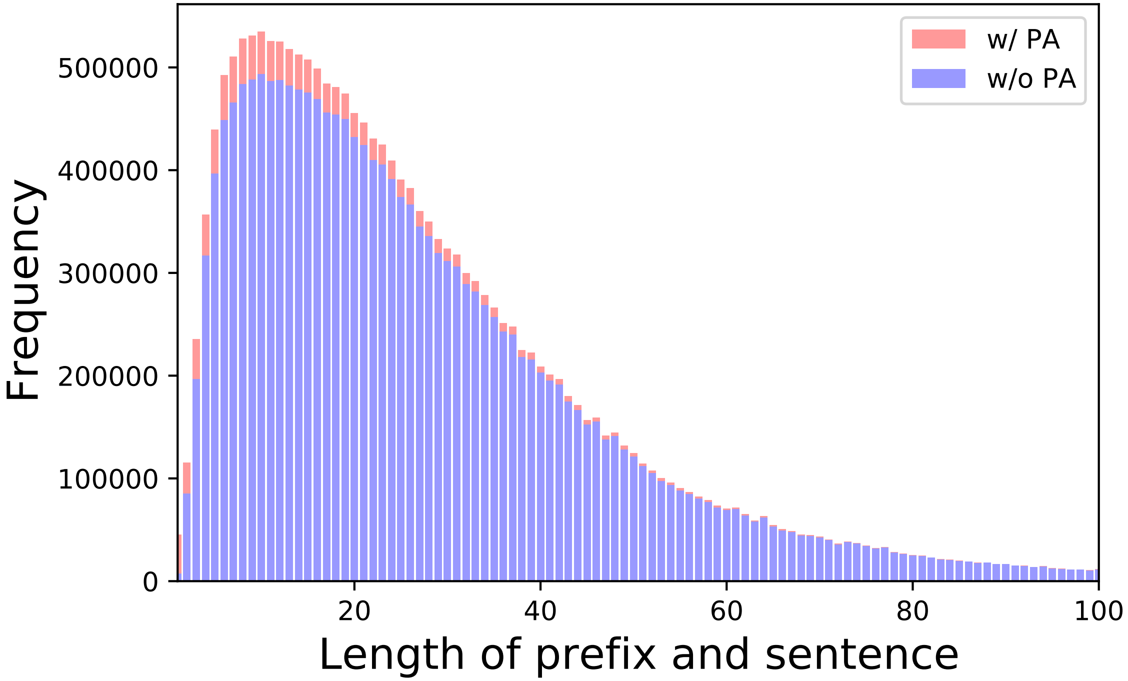


Source sentence length distribution (train data)

En-De



En-Ja



Number of extracted prefixes are different because of word order difference

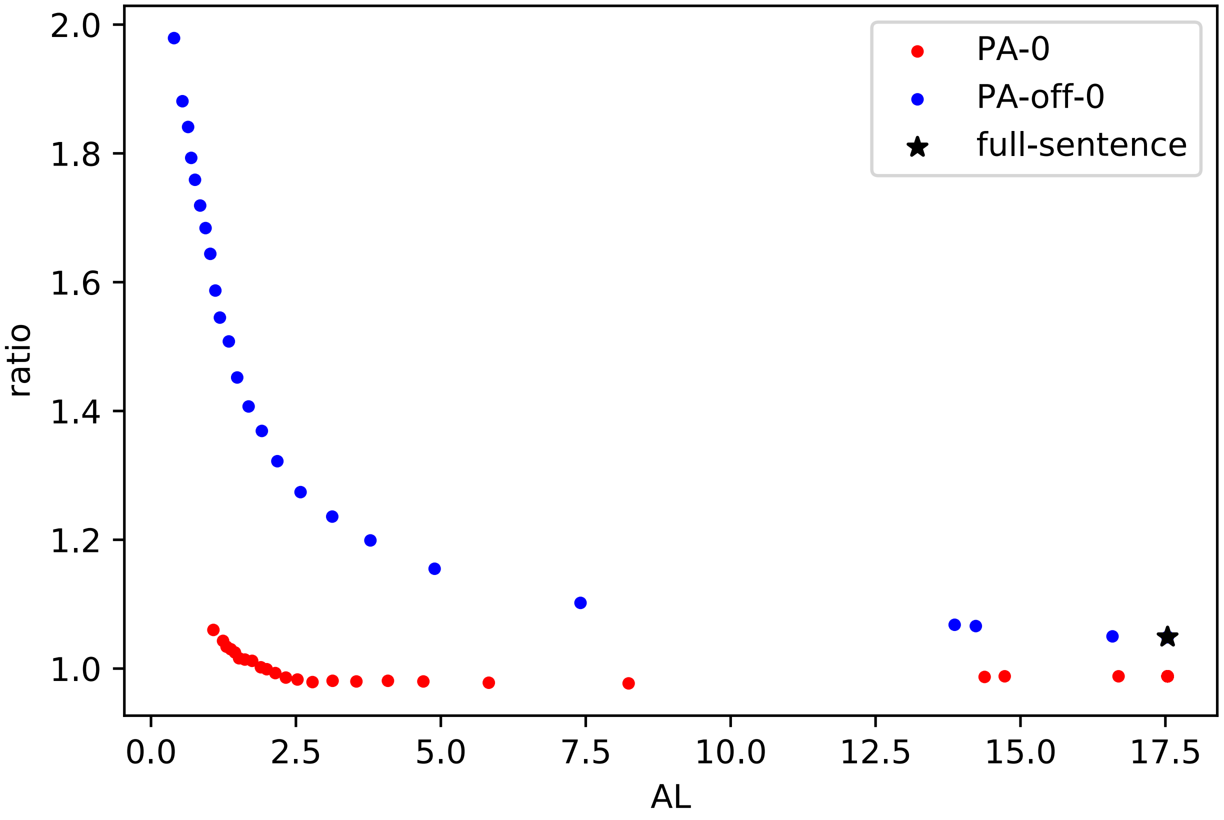
Conclusion

- Proposed method: Fine-tune NMT model with bilingual prefix pairs for simultaneous translation
  - Decreased length ratio
  - Outperformed baselines in quality-latency trade-off in low latency

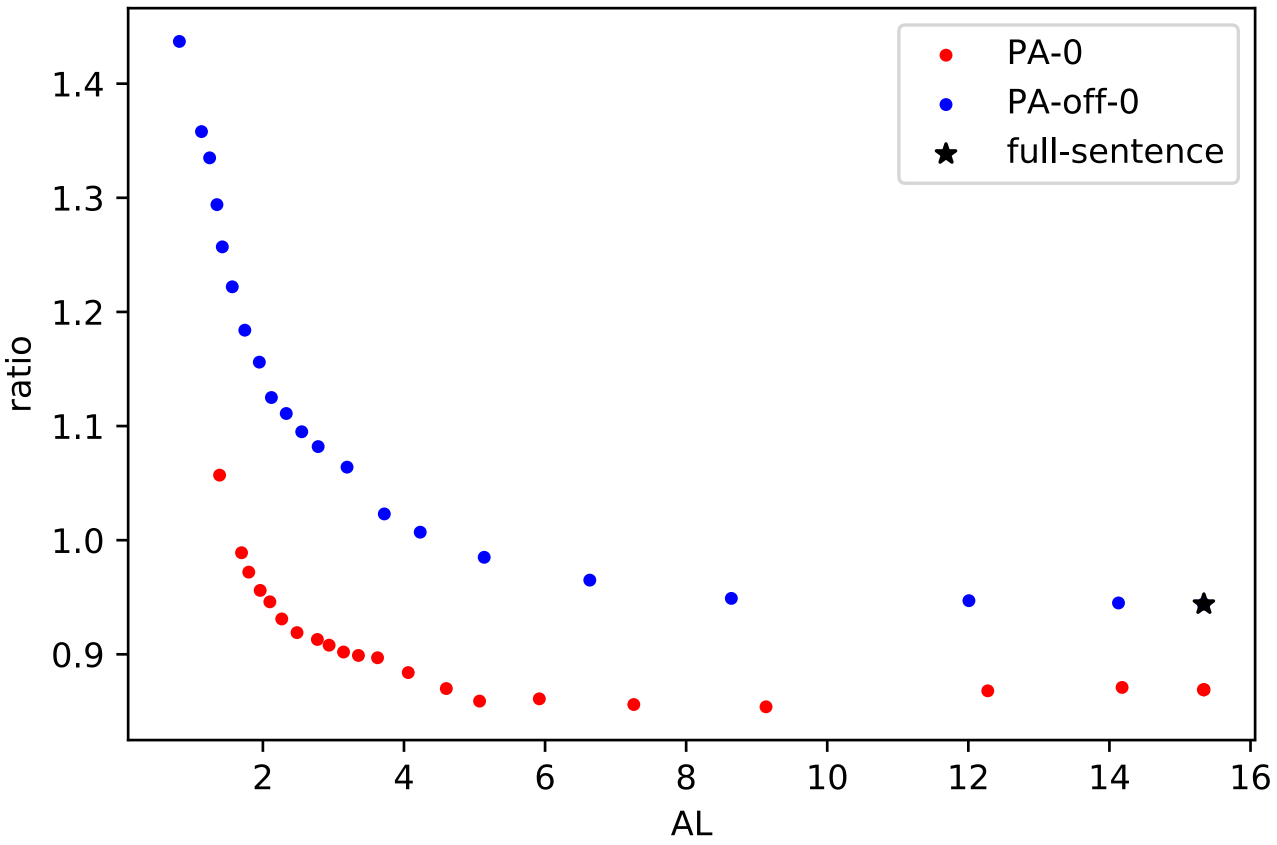
- Future work
  - Work for language pairs with different word order
  - End2end Speech 2 text (implemented as the system submitted to IWSLT 2022 Evaluation Campaign)

Effect of Fine-tuning (Length Ratio)

En-De



En-Ja



Memo

- Compared with word alignment
  - The proposed method find boundary which is suitable for the pretrained NMT model to translate. (segmentation by Word Alignment is separated from the training of NMT model.)
  - Easily applied to end2end speech translation

Reference

[Vaswani+, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

[Zhang+, 2020] Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2020. Learning adaptive segmentation policy for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2280–2289, Online. Association for Computational Linguistics.

Reference

[Devlin+,2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language under- standing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Reference

[Neubig+, 2014]Graham Neubig, Katsuhito Sudoh, Yusuke Oda, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. 2014. The NAIST-NTT TED talk treebank. In *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, USA.

[Murcus+, 1993]Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Reference

- [Ma+ , 2019] Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.