Simultaneous Neural Machine Translation with Constituent Label Prediction

Yasumasa Kano[1], Katsuhito Sudoh[1,2], Satoshi Nakamura[1,2]

1.Nara Institute of Science and Technology (NAIST), Japan

2.Center for Advanced Intelligence Project (AIP), RIKEN, Japan

Simultaneous translation

- Consecutive translation

  Input: I am a student .

  Output: 私 は 学生 です 。

- Simultaneous translation

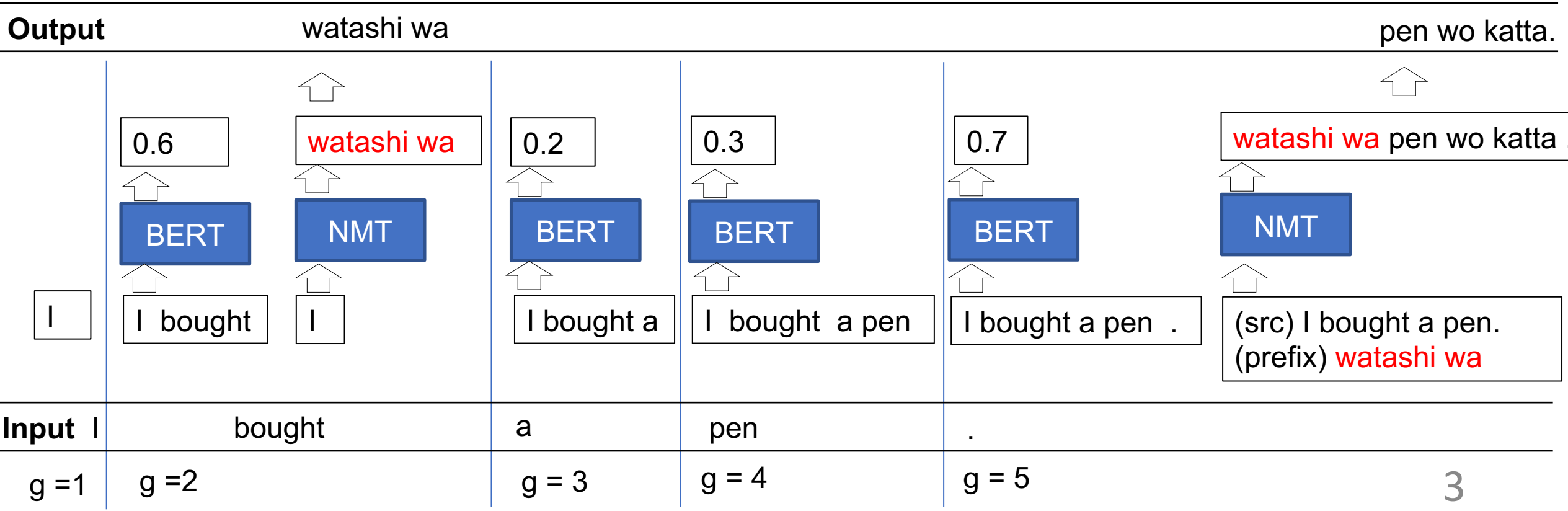
  Input: I am a student .

  Output: 私 は 学生 です 。

Related work: Meaningful Unit [Zhang+, 2020]

Threshold: 0.5

Future words: 1



Language pairs with different word orders

- SVO(Subject-Verb-Object) → SOV

  En) I bought a pen .

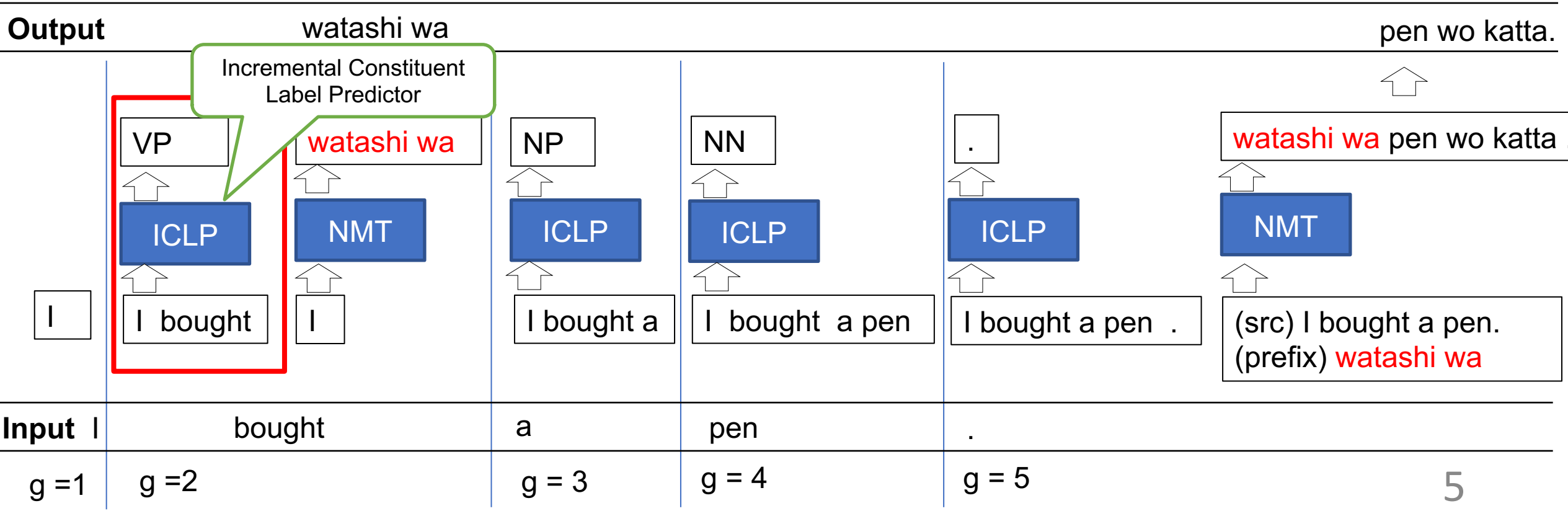
  Ja) Watashi wa pen wo katta .

  → Use syntax information

Proposed: Constituent Label Prediction

Result of full-sentence parser (S (NP (PRP I)) (VP (VBD bought) (NP (DT a) (NN pen))))

Future words: 1



Simple rules based on predicted labels

- Segment the input coming just before constituents labeled S and VP.

I / saved time by / doing this.
VP NP PP S NP

- If the previous label is S or VP, do not segment the input.

I / can (/) save time .
VP VP NP

- If the chunk is shorter than the minimum length, do not segment the input. [Change minimum length to adjust latency]

I (/) bought a pen .
VP NP NN

Minimum length = 2

Experiment of ICLP

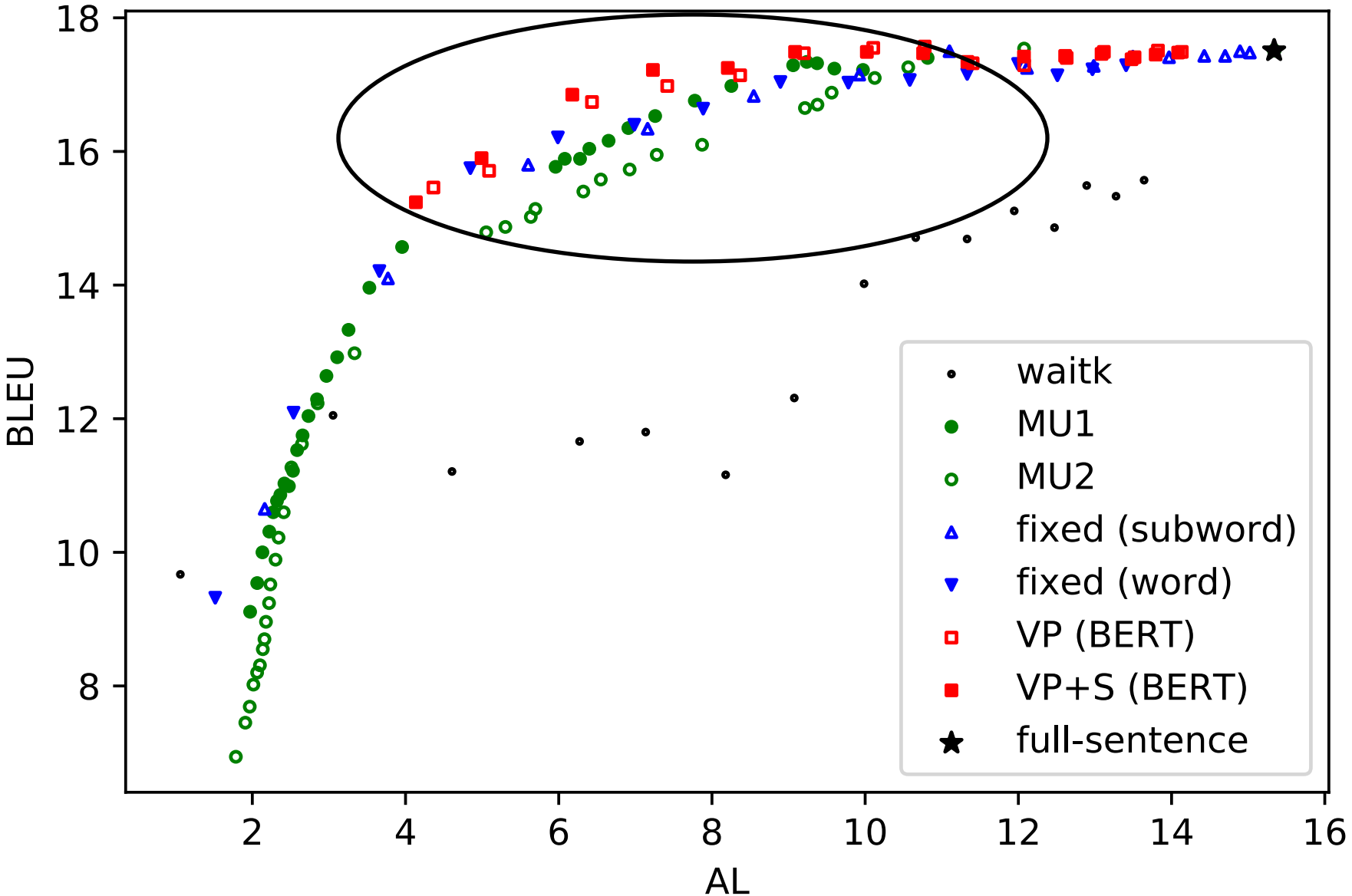
- Data
  - Train: Penn Treebank 3 [Marcus+, 1993]
  - dev: 1% of training data
  - test: NAIST-NTT TED Talk Treebank [Neubig+, 2014]
- Model
  - BERT [Devlin+,2019]
- Result (VP)

Model	Precision	Recall
0 future words	0.75	0.80
1 future word	0.89	0.97
1 future word (LSTM)	0.91	0.94

Experiment of simultaneous translation

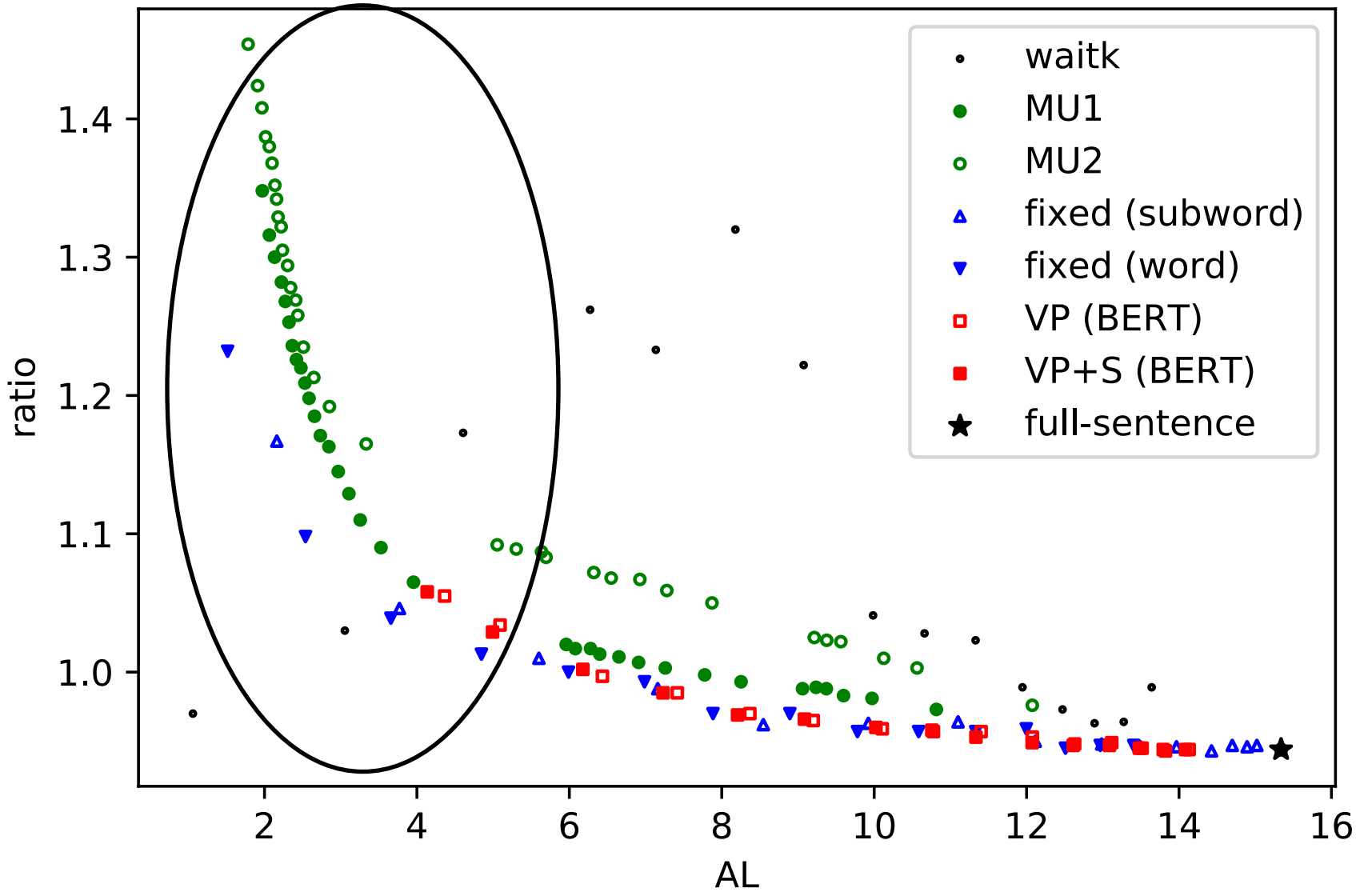
- Data [En-Ja]
  - Pretrain: 20M (WMT2020
  - Fin-tune: 200K (IWSLT2021
  - Dev: 5.3K (IWSLT dev2010, tst2011, tst2012, and tst2013
  - Test: 1.5K (IWSLT2021 dev
- Subwords
  - Joint vocabulary size 16k (BPE)
- NMT Model
  - Transformer [Vaswani+, 2017]
- Evaluation metrics
  - Quality: BLEU
  - Latency: AL (Average Lagging) [Ma+ , 2019]

Result [En→Ja]



Proposed method outperformed baselines for wide range of AL.

Length Ratio [En→Ja]



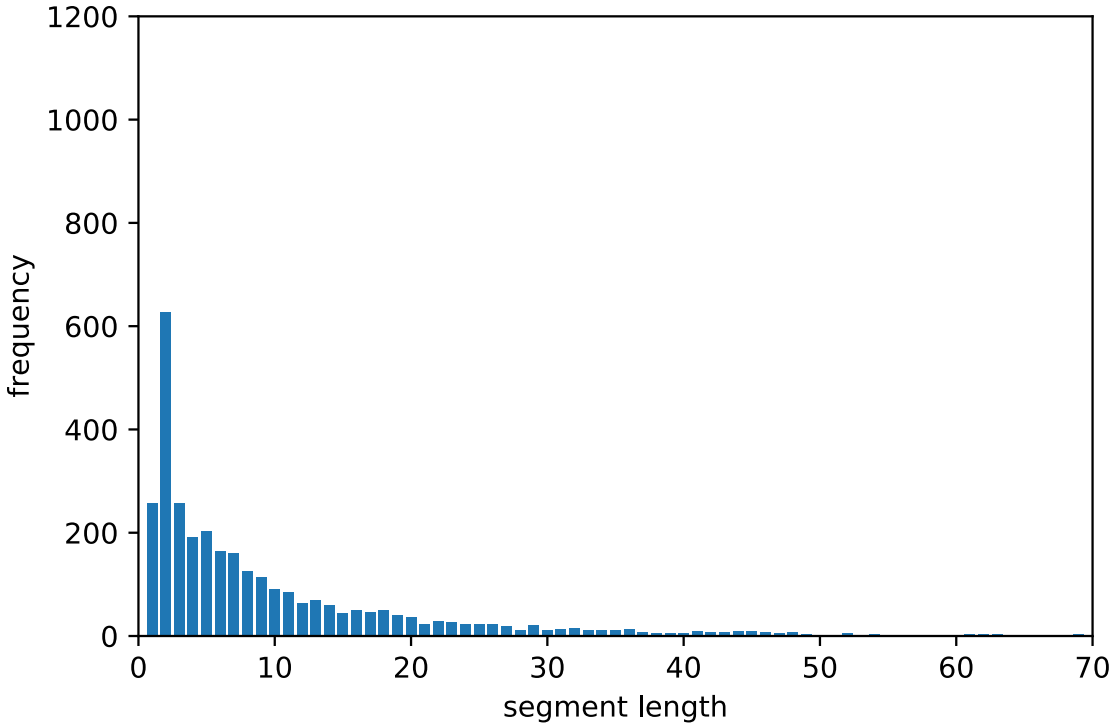
Translation of short segments tends to be longer than expected

Segment Length Distribution (test [En→Ja])

Meaningful Unit (1 future word)

AL: 7.26

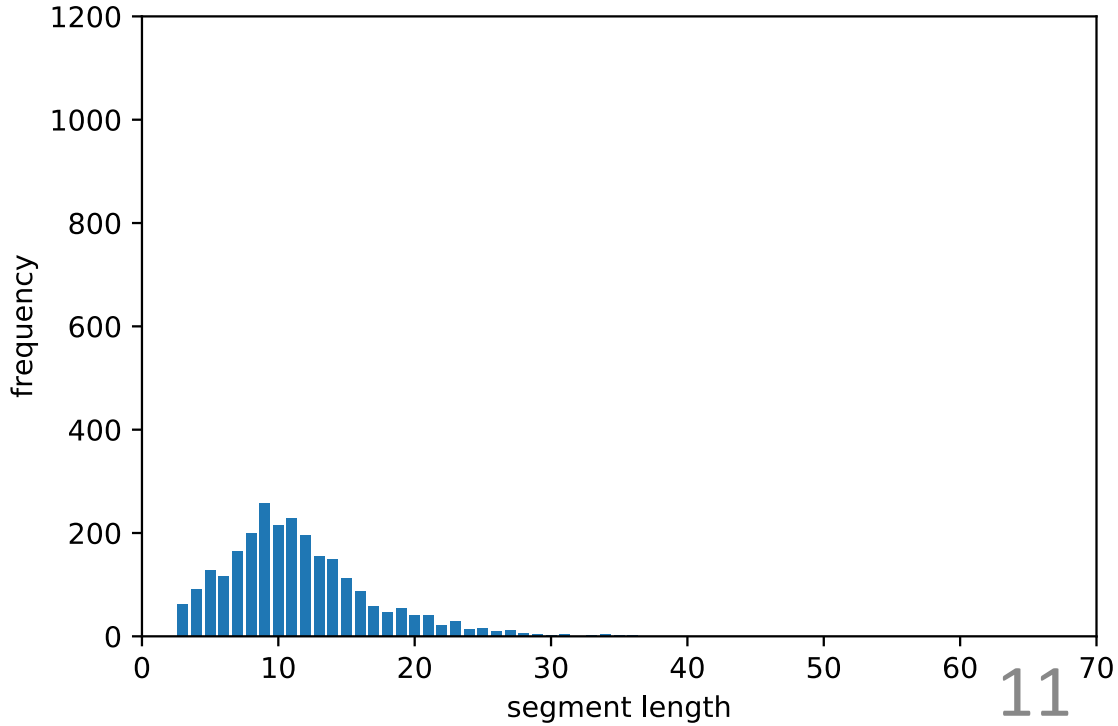
BLEU: 16.53



Proposed: ICLP (1 future word)

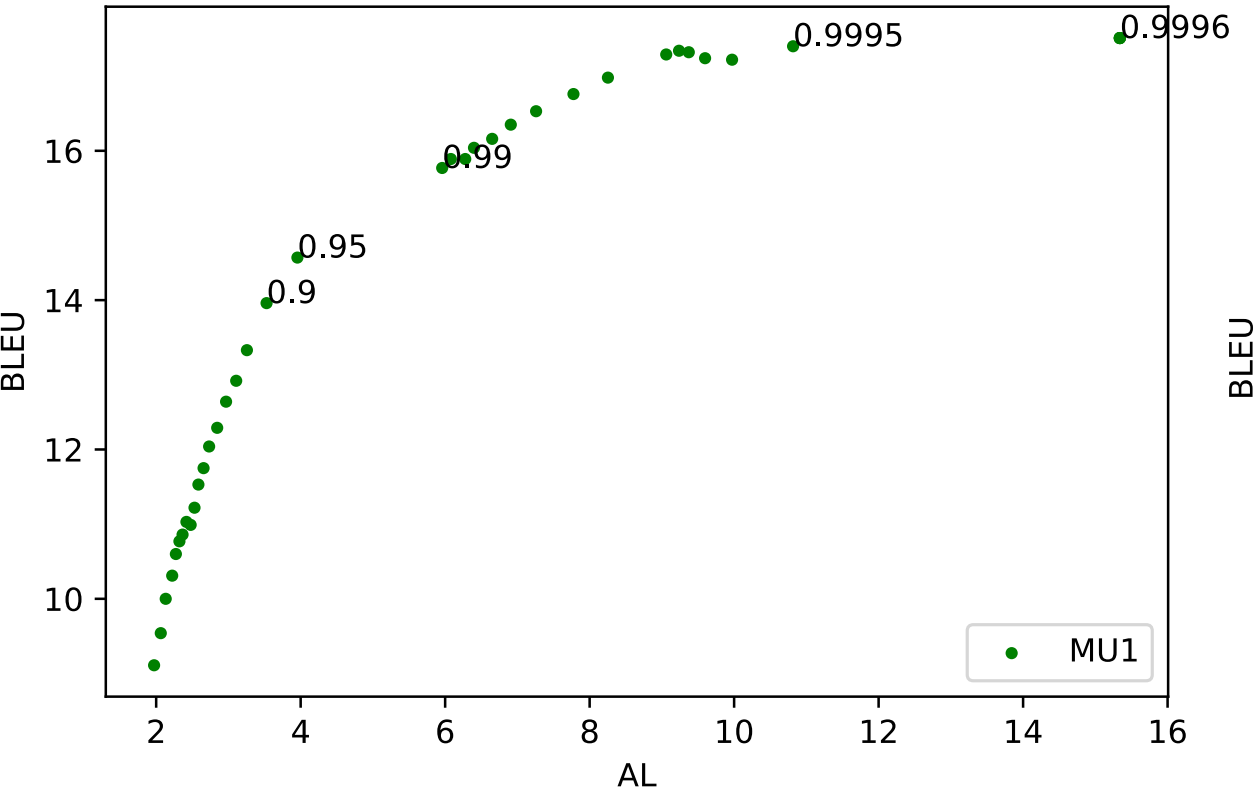
AL: 7.23

BLEU:17.22

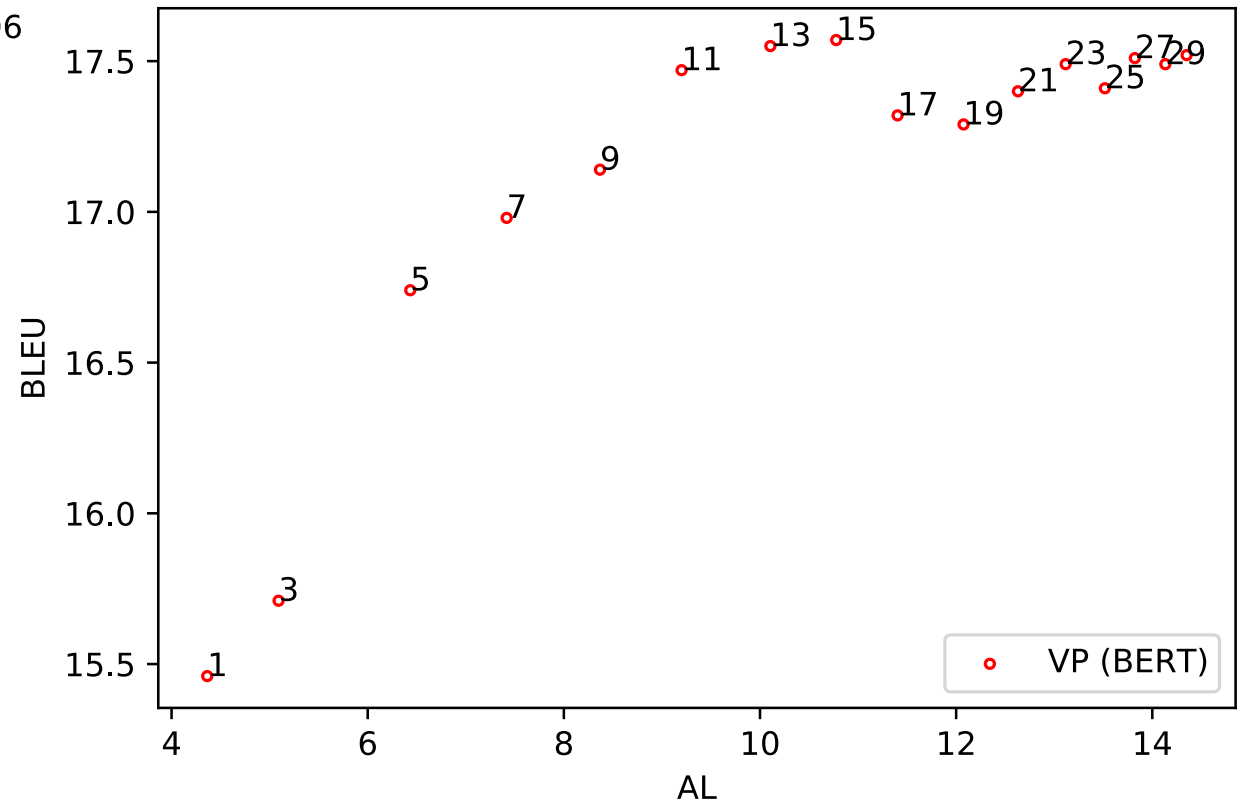


Controllability of latency [En→Ja]

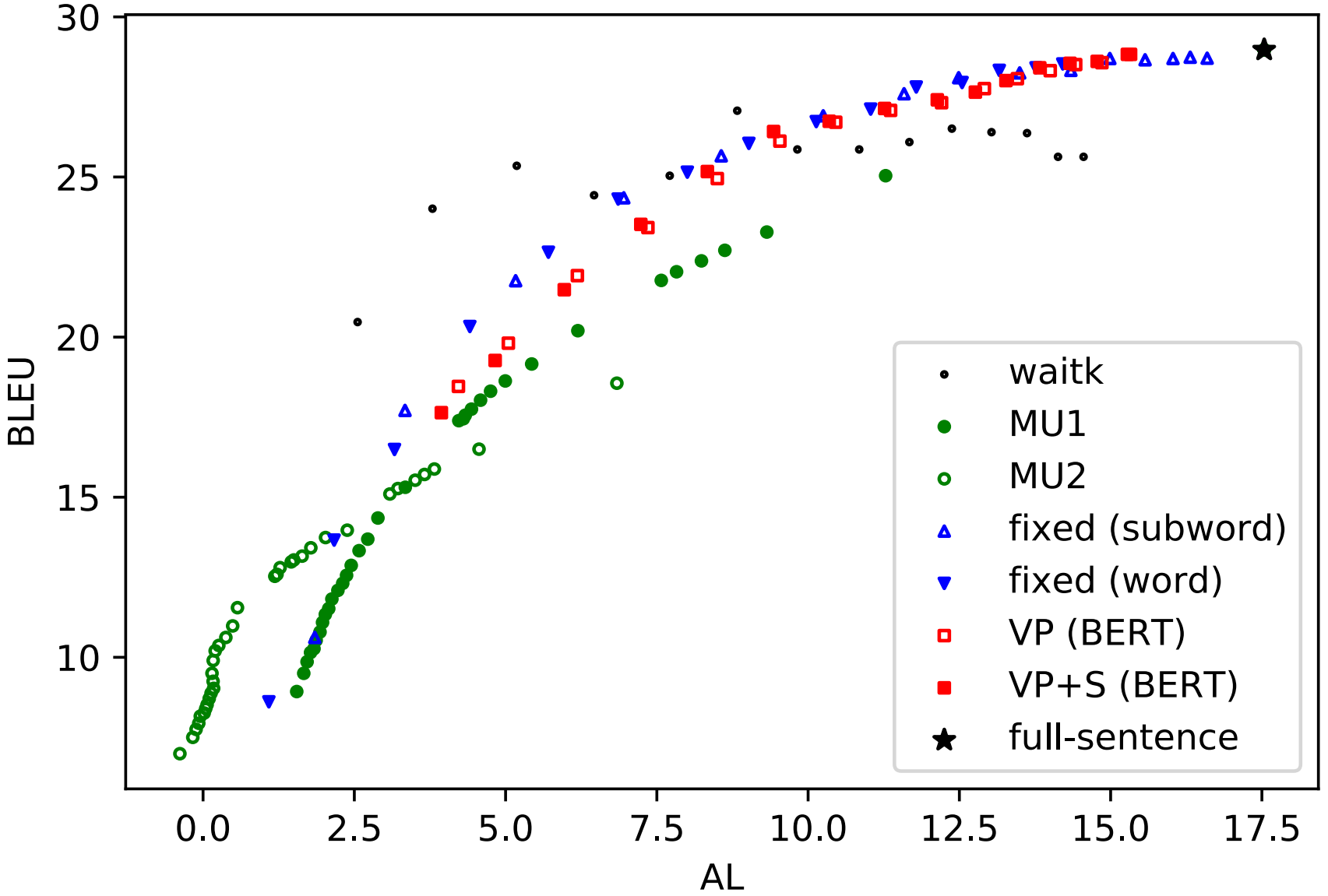
Meaningful Unit



Proposed: ICLP



Result [En→De]



Look-ahead approaches did not improve the performance.

Conclusion

- Novel segmentation method: simple rules and label predictor
  - Higher BLEU than baselines in En-Ja simultaneous translation
  - Easy to control latency
  - Not dependent on trained NMT model

- Future work
  - Extract rules automatically
  - Improvement for other language pairs

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.