

構文ラベル予測による 同時ニューラル機械翻訳

言語処理学会2022 C6-2

加納保昌¹, 須藤克仁^{1,2}, 中村哲^{1,2}

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

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

同時翻訳

- フルセンテンス翻訳

入力: I am a student .

出力: 私 は 学生 です 。

- 同時翻訳

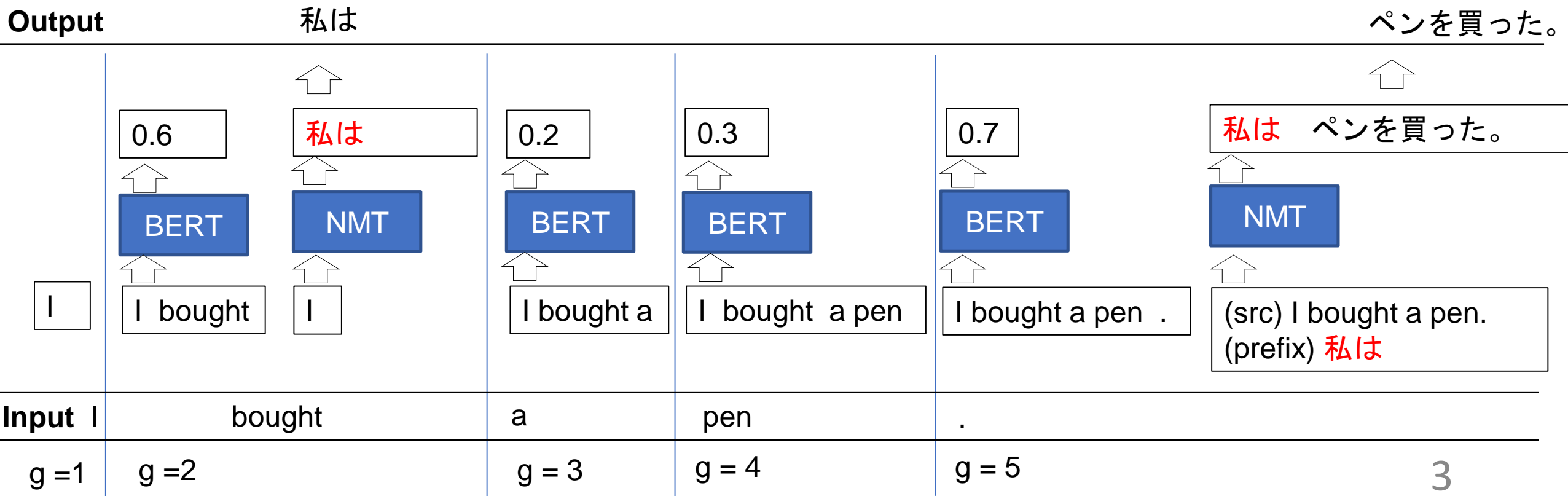
入力: I am a student .

出力: 私 は 学生 です 。

先行研究: Meaningful Unit [Zhang+, 2020]

Threshold: 0.5

Future words: 1



語順の異なる言語

- SVO(Subject-Verb-Object) → SOV

En) I bought a pen .

Ja) 私はペンを買った .

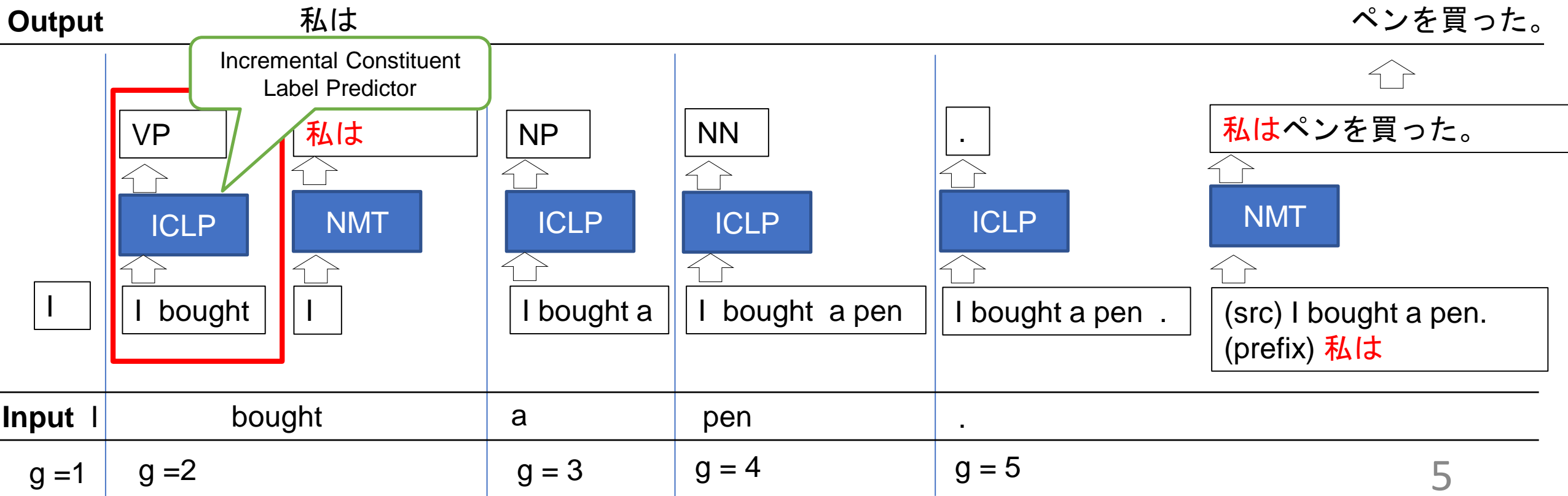
→ 構文情報を用いる。

提案手法: 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

- S と VP の直前で区切る。

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

- 前のラベルが S または VP であれば区切らない。

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

- チャンクが最小セグメント長より小さければ区切らない

[**最小セグメント長**で遅延を調整]

I (/) bought a pen .
VP NP NN

最小セグメント長 = 2

ICLPの実験

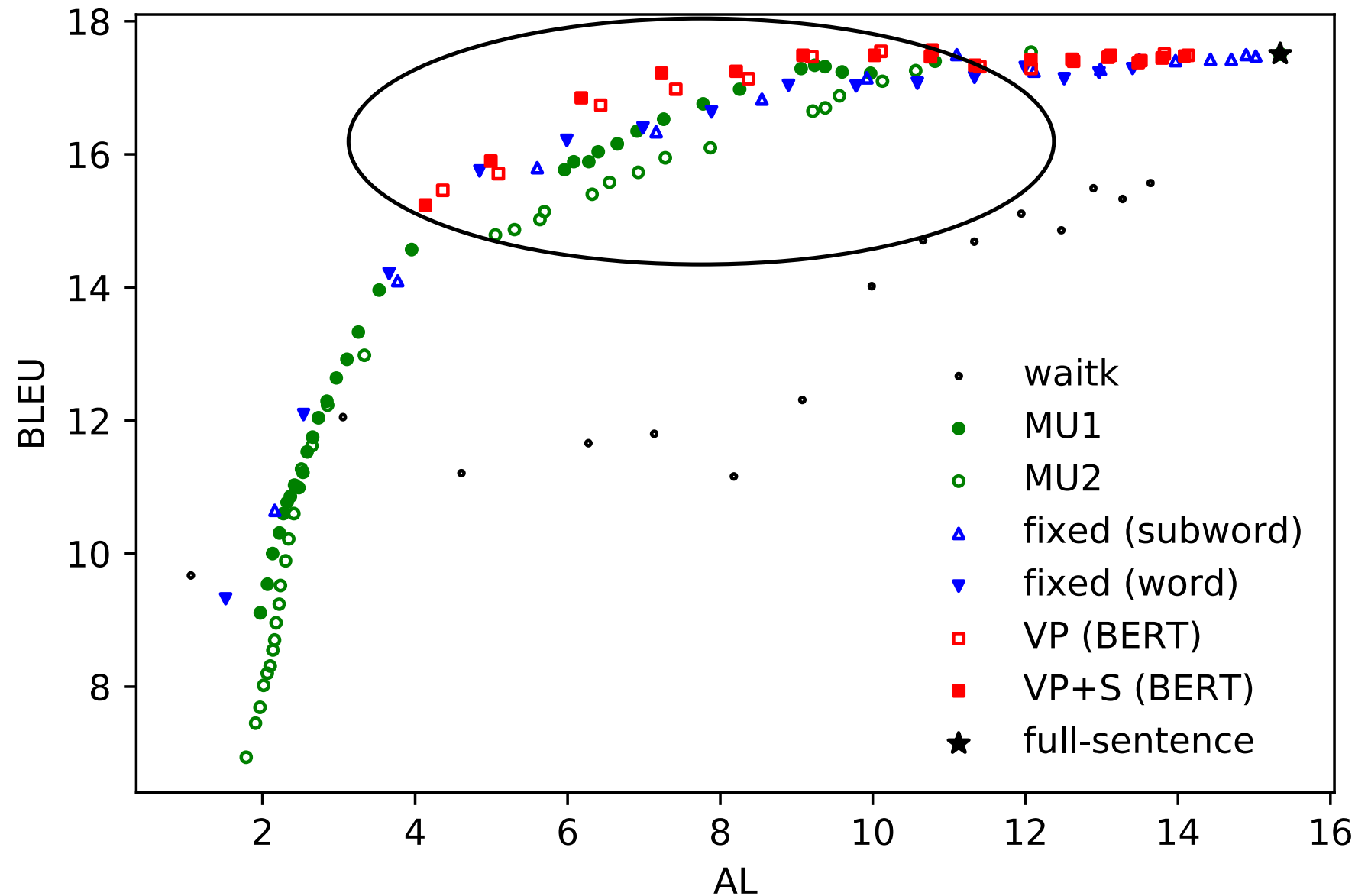
- データ
 - Train: Penn Treebank 3 [Marcus+, 1993]
 - Dev: 1% of training data
 - Test: NAIST-NTT TED Talk Treebank [Neubig+, 2014]
- モデル
 - BERT [Devlin+,2019]
 - LSTM
- 結果 (VP)

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

同時翻訳の実験

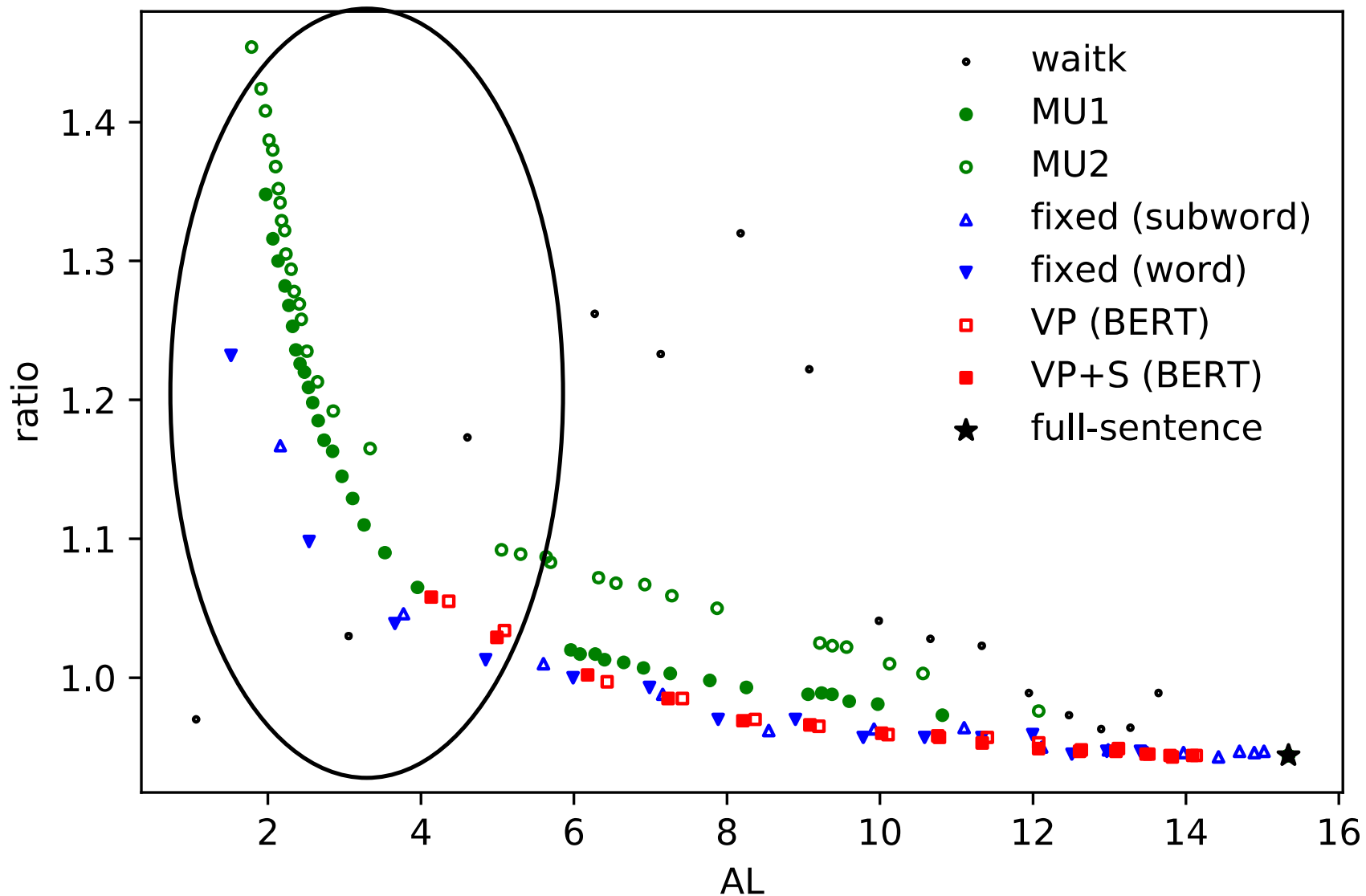
- データ [En-Ja]
 - Pretrain: 20M (WMT2020)
 - Fin-tune: 200K (IWSLT2021)
 - Dev: 5.3K (IWSLT dev2010, tst2011, tst2012, and tst2013)
 - Test: 1.5K (IWSLT2021 dev)
- サブワード
 - Joint vocabulary size 16k (BPE)
- NMT モデル
 - Transformer [Vaswani+, 2017]
- 評価指標
 - 精度: BLEU
 - 遅延: AL (Average Lagging) [Ma+ , 2019]

結果



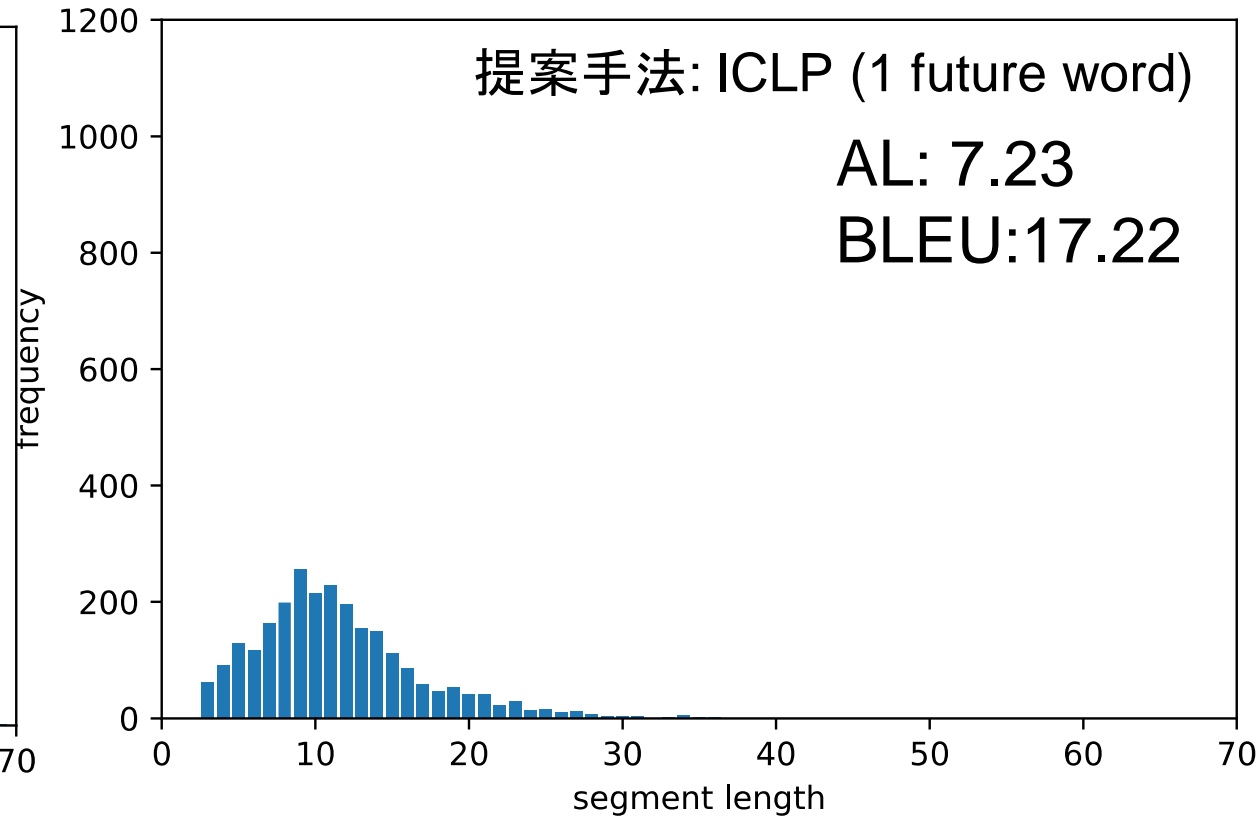
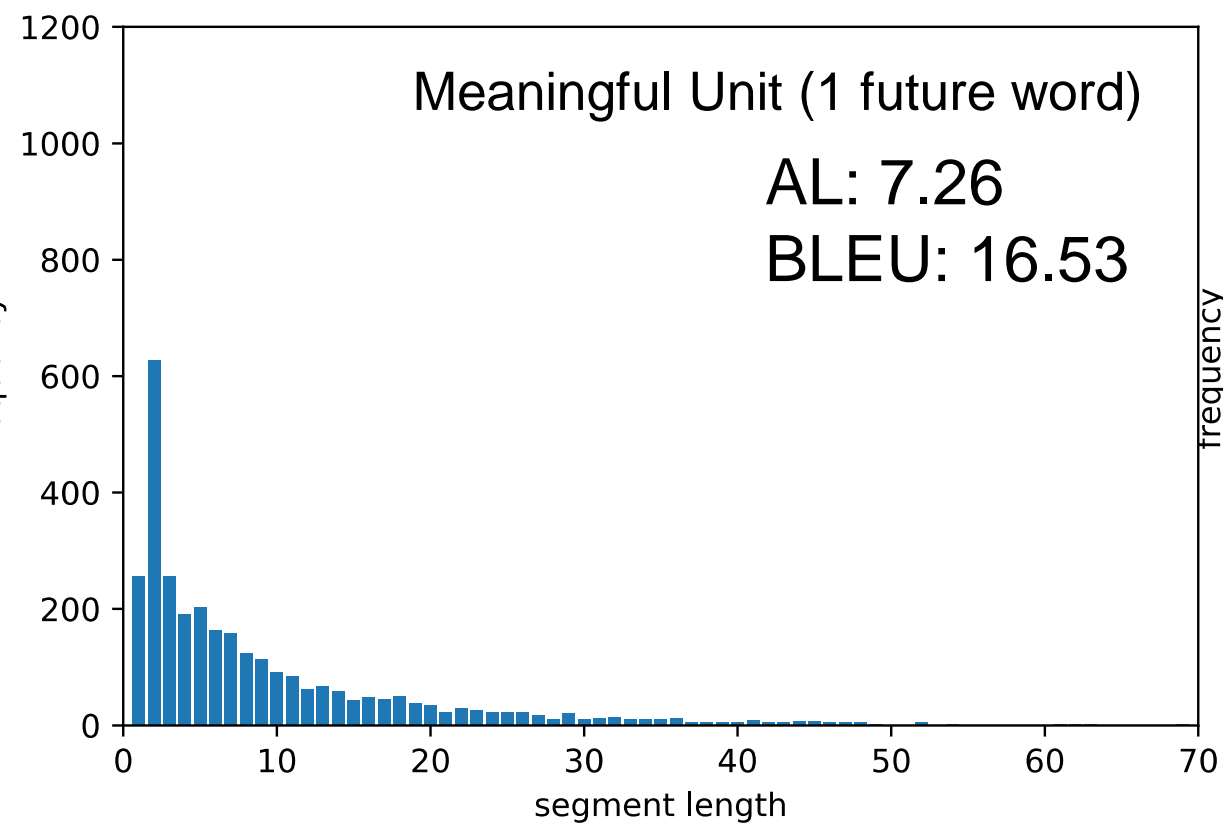
提案手法は精度と遅延のトレードオフでベースラインを上回った。

Length Ratio



短いセグメントの翻訳は過剰に長くなりがち

Sourceセグメント長分布 (Test data)



- Meaningful Unitは短いセグメントが多く、ratioを大きくした。
- 平均の遅延は同程度だが、Meaningful Unitは分散が大きく遅延の制御がしづらい

構文の制御 (出力例)

原文 I /like / delicious food .

ラベル / VP / NP / NN / .

翻訳 私は / 好きです / 美味しい食べ物.

順送り調

VP+NP で区切った結果

原文 I / like delicious food .

ラベル / VP / NP / NN / .

翻訳 私は / 美味しい食べ物が好きです.

フルセンテンス調

VP のみで区切った結果

まとめ

- セグメンテーション方法を提案
 - 構文ラベル予測モデルとシンプルなルール
- 結果
 - 英日同時翻訳タスクでベースラインを遅延と精度のトレードオフで上回った。
 - 遅延や出力構文の制御がしやすくなった。
- 今後
 - 別の言語対にも適用できるようにする。

Average Lagging: Wait-k

K = 2

SRC SRC

SRC

SRC

TGT

TGT

TGT TGT

$g(t) : [2,3,4]$ (t番目のTGTを読む前に何個のSRCを読んだか。)

$t : [1,2,3]$ (t番目のTGT)

$t - 1 : [0,1,2]$ (t - 1)

AL: $\text{sum}(g(t) - (t-1)) / \text{len}([0,1,2]) = ((2-0)+(3-1)+(4-2))/3 = 2$

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 understanding. 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.