NAIST Simultaneous Speech-to-Text Translation System for IWSLT 2022

Ryo Fukuda[1], Yuka Ko[1], Yasumasa Kano[1], Kosuke Doi[1], Hirotaka Tokuyama[1], Sakriani Sakti[1,2], Katsuhito Sudoh[1], and Satoshi Nakamura[1]

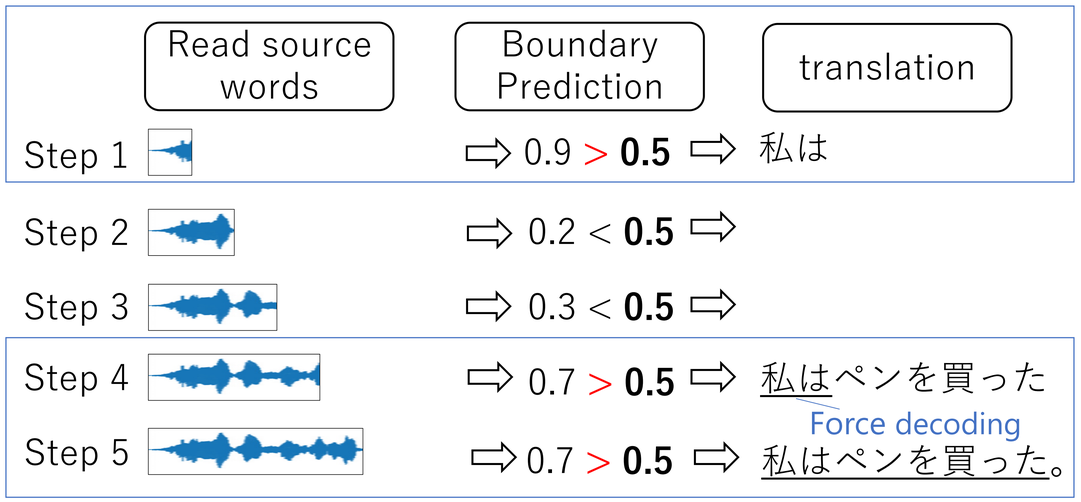
[1]Nara Institute of Science and Technology (NAIST), Japan [2]Japan Advanced Institute of Science and Technology (JAIST), Japan

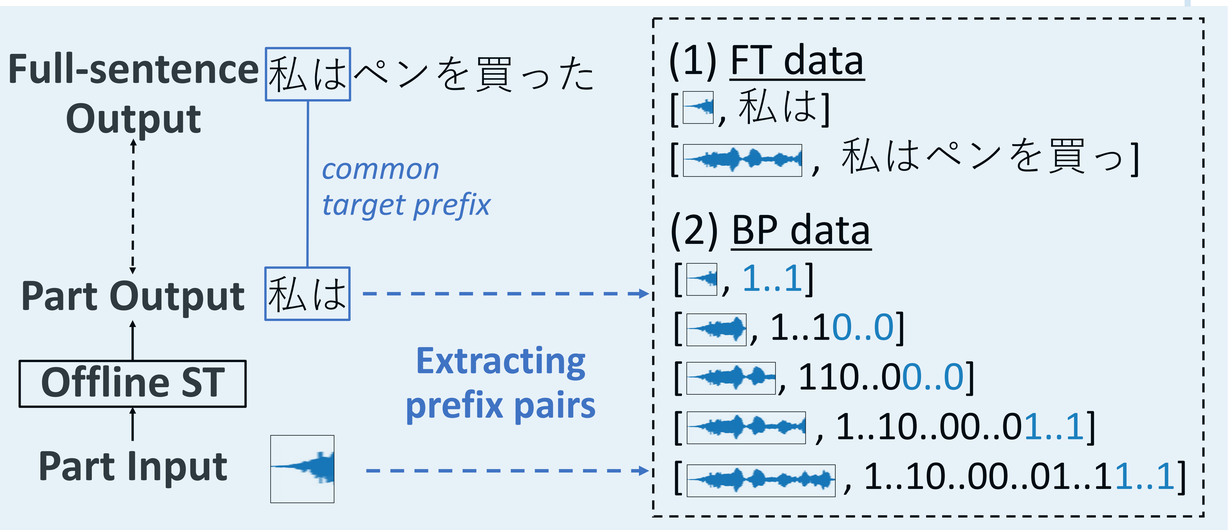
Abstract End-to-end SimulST using adaptive segmentation policies based on bilingual prefix alignment [Kano et al., 2022]

Bilingual Prefix Alignment (BPA) for SimulST

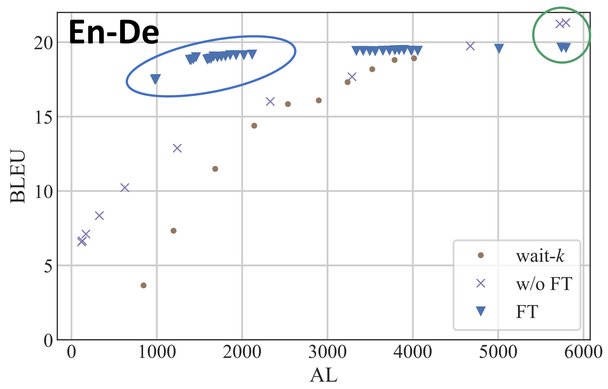
- Training process
  - (1) Fine-tune (FT) offline ST
  - (2) Boundary Predictor (BP)

  } from bilingual prefix pairs data
- Translation process
  - BP detects segment boundaries
  - ST translates a partial input taking translation history into account

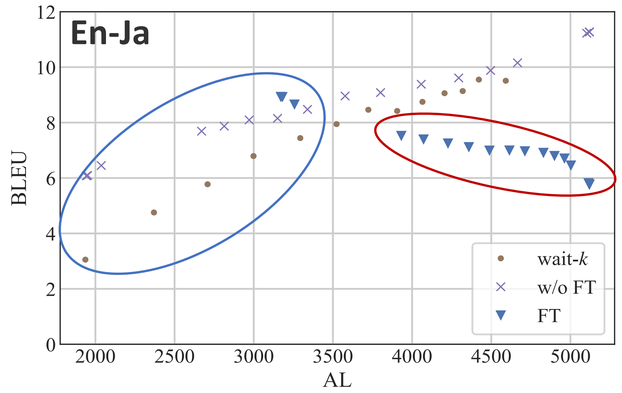




Experiment Data MuST-C v2 En-De, En-Ja **ST Model** Transformer
Boundary Prediction Model LSTM, 100 frames/unit
Evaluation MuST-C v2 tst-COMMON in SimulEval



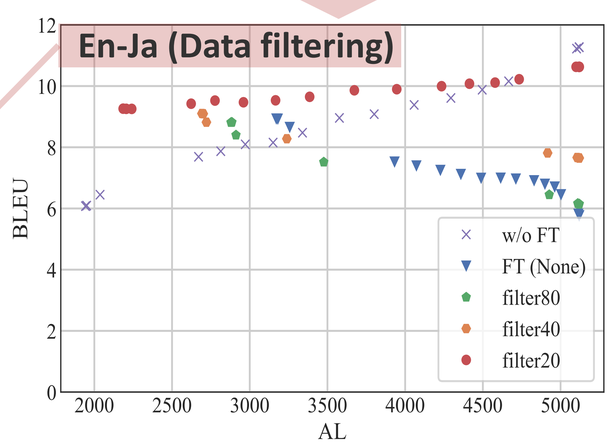
- FT > w/o FT in AL<=4000
  → Robust to lower latency
- w/o FT > FT in high latency



- w/o FT > wait-*k*
- FT < wait-*k*, w/o FT
  → prefer too short outputs

Removing prefix pairs from FT data if len(src)/len(tgt) > *maxratio*

- Unbalanced prefix pairs would cause degradation
  - ➢ e.g. {En, Ja} prefix pair would consist of {S,S}, {SV, S}, {SVO, SOV}



- FT with filtered data achieved best performance!

Conclusion Our results show effectiveness of fine-tuned ST & Boundary Predictor with bilingual prefix alignment

- ➢ Robust to lower latency
- ➢ En-Ja : data filtering of large length gap pairs was effective due to its sentence structure