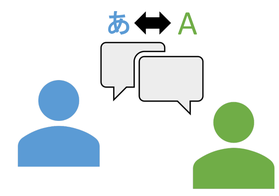
Towards Speech Dialogue Translation Mediating Speakers of Different Languages

Shuichiro Shimizu[1] Chenhui Chu[1] Sheng Li[2] Sadao Kurohashi[1,3]
[1]Kyoto University
[2]National Institute of Information and Communications Technology, Japan
[3]National Institute of Informatics, Japan

1. Background

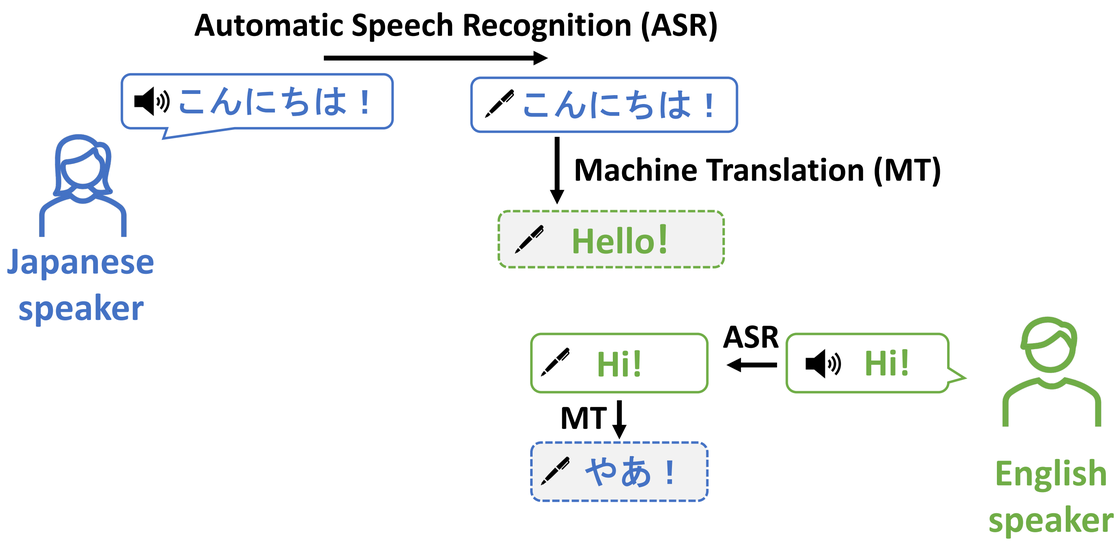
- Communication across language boundaries is becoming increasingly important in this global era
- Current speech translation research mostly focuses on monologue



Speech translation focusing on **cross-language dialogue** is needed

2. Speech Dialogue Translation (SDT)

- Translate each utterance from the speaker's language into the listener's language
- We consider cascade of ASR and MT in this study



3. SpeechBSD Dataset

- No dataset for SDT
  → Crowdsource audio of existing dialogue MT corpus
- Business Scene Dialogue (BSD) corpus (Rikters et al., 2019)
  - Manually designed business scene dialogues
  - Parallel corpus of English and Japanese
    → Regard as cross-language dialogue

Crowdsourcing Platform

Ja YAHOO! JAPAN
En amazon mechanical turk

Collect speaker attributes (gender, homeplace)

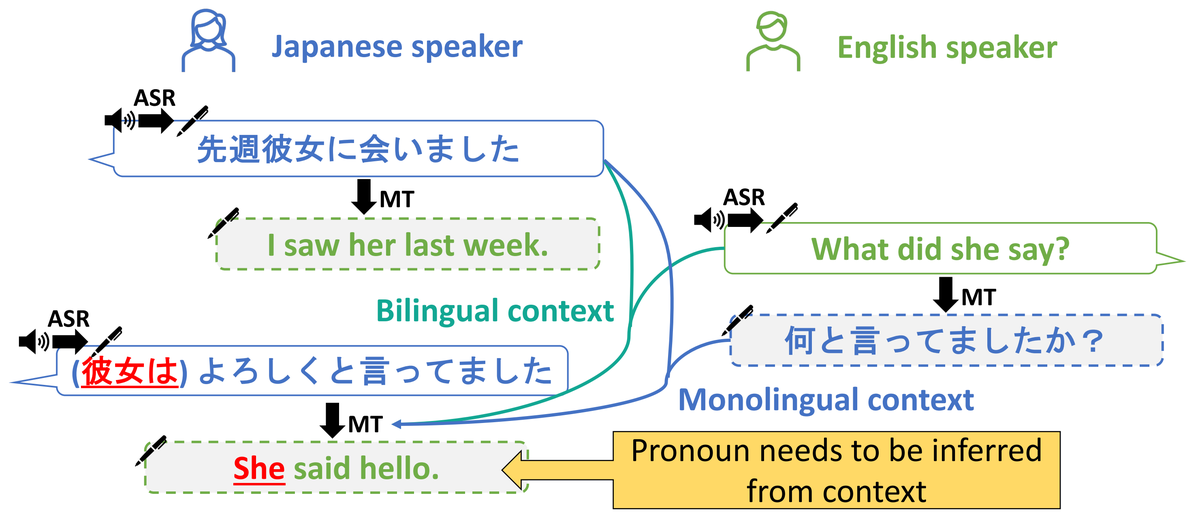
Audio Collection Webpage

Record & check utterances

Statistics

	Train	Dev.	Test
Scenarios	670	69	69
Sentences	20,000	2,051	2,120
En audio (h)	20.1	2.1	2.1
Ja audio (h)	25.3	2.7	2.7

4. Considering Context in SDT

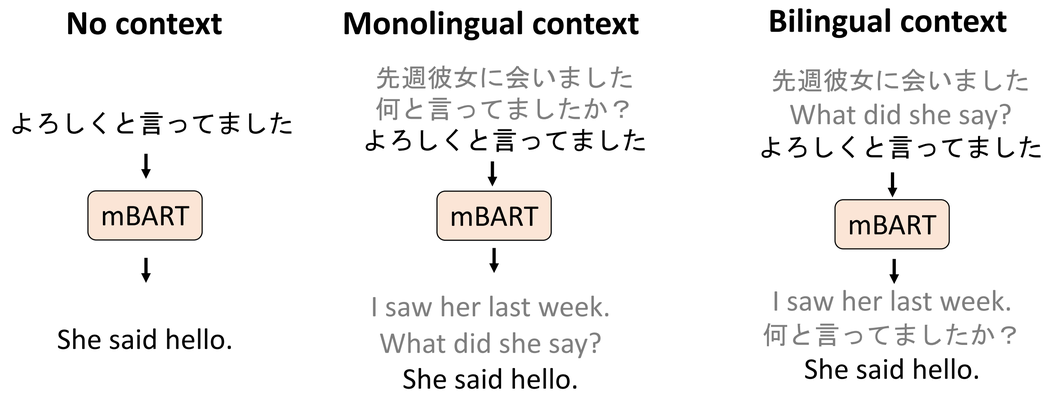


5. Experiments

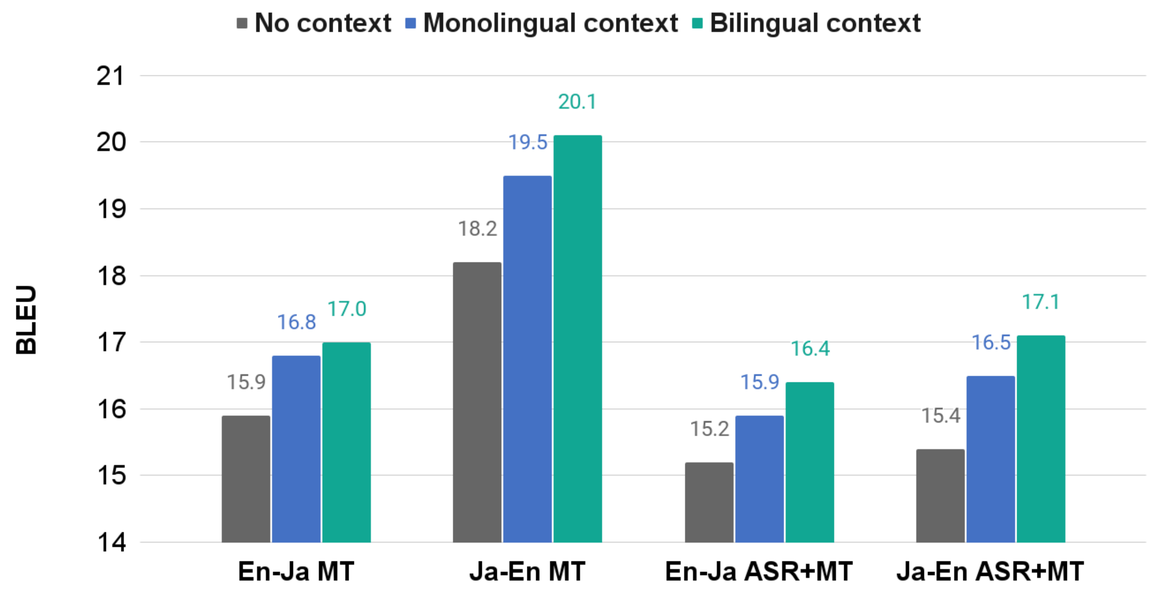
Models

- ASR: Whisper (Radford et al., 2022)
  - Multilingual ASR model
  - Robust performance with 680,000h training data
- MT: mBART (Liu et al., 2022)
  - Self-supervised learning in 25 languages
  - MT becomes possible with finetuning

Settings



Results



6. Conclusion

- Proposed speech dialogue translation focusing on cross-language dialogue
- Constructed SpeechBSD dataset for the task
- Showed considering context in two languages performs well

Future Work

- End-to-end speech translation
- Speech-to-speech translation considering speaker attributes