Recognition and Translation of Code-switching Speech Utterances

Sahoko Nakayama[1,2], Takatomo Kano[1], Andros Tjandra[1], Sakriani Sakti[1,2], Satoshi Nakamura[1,2]

[1]Nara Institute of Science and Technology, Japan

[2]RIKEN, Center for Advanced Intelligence Project AIP, Japan

{nakayama.sahoko.nq1, kano.takatomo.km0, andros.tjandra.ai6, ssakti, s-nakamura}@is.naist.jp

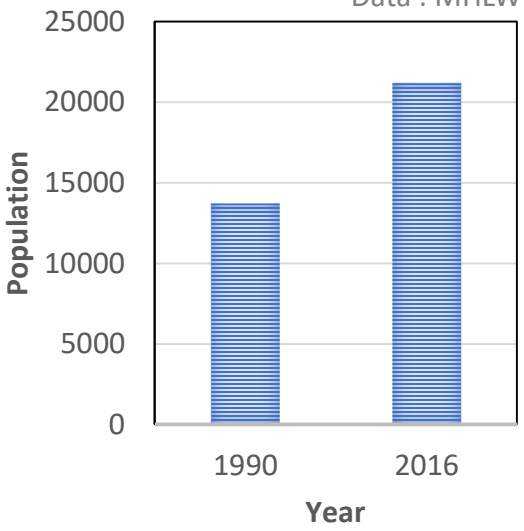
Background

Bilingualism

Bilingual speakers have increased in Japan

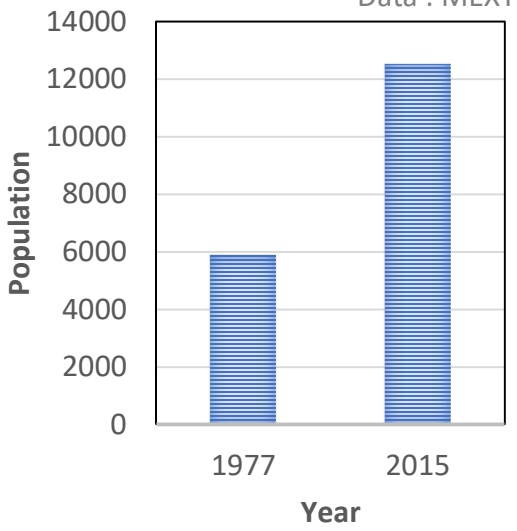
Children with a foreign parent

Data : MHLW



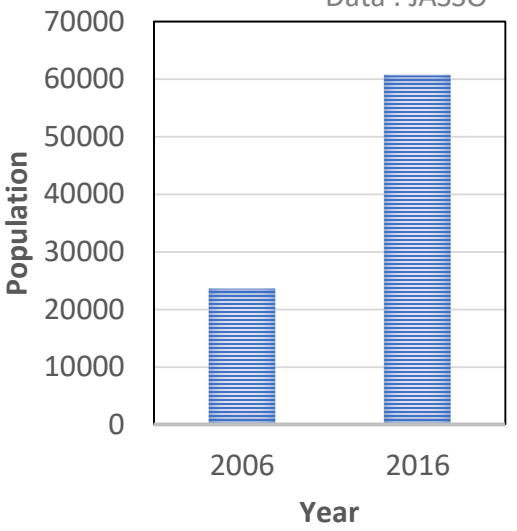
Children who have lived abroad

Data : MEXT



Students who studied abroad

Data : JASSO



Code-switching (CS) plays an important role in bilingualism [McSwan, 2000].

CS Definition

Speakers switch languages within a conversation.

- Word-level CS:

国会が the Equal Employment Opportunity Law に罰則を設けなかったので、空文だという意見があります。

(*As the Diet did not put any teeth into the Equal Employment Opportunity Law, some are of the opinion that it is a mere scrap of paper.)*

- Phrase-level CS:

If I could make a suggestion, この議題についての討議を昼食までに終えて頂ければと思いますが。

(If I make a suggestion, would you finish discussing this subject by lunch time?)

CS Coverage of This Study

The definition of what constitutes CS is controversial.

- Are loanwords word-level CS?

  中間言語を使った時のメリットに何があるか？

  (*What is the merit of using an interlingua?*)

- Are quotations phrase-level CS?

  What do you think of the Japanese saying, "うそつきは泥棒の始まり"?

  (*What do you think of the Japanese saying, "Show me a liar and I'll show you a thief"?*)

Theoretically, they may not be CS.

But in this study we aim to properly handle as many cases as possible.

Therefore, we will try to handle these cases as well.

How Code Switching Occurs

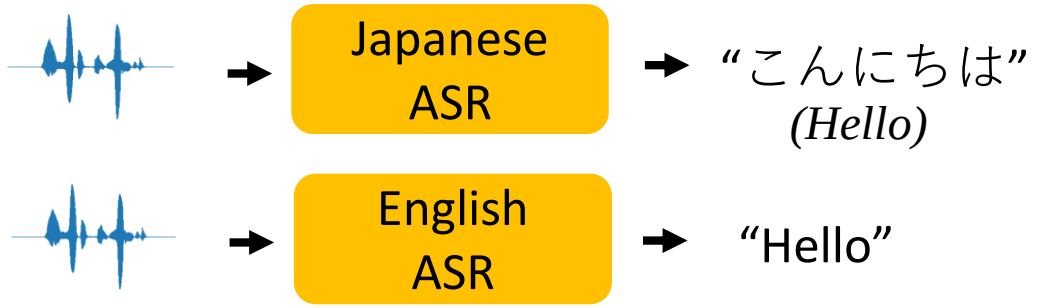
- Proficiency-driven CS
  - A speaker is competent in both languages
  - Easily able to switch from one language to another.

- Deficiency-driven CS
  - A speaker is lack of competency of one language
  - Go back to another language.

As the data of deficiency-driven CS has not been obtained yet, we only handle the proficiency-driven case in this work.

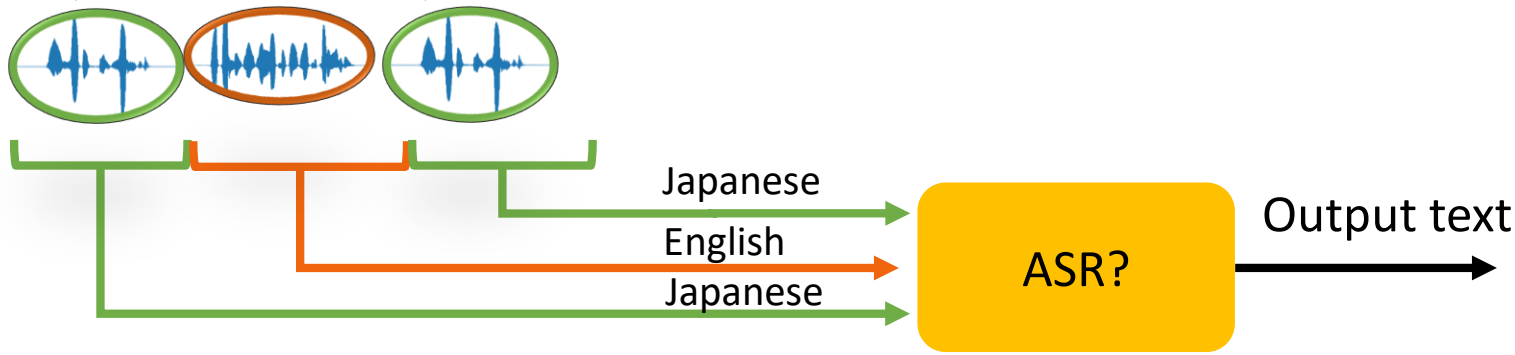
CS Challenges For ASR

- Standard Automatic Speech Recognition (ASR) is monolingual



- Challenge with CS: need to handle multilingual input

“これはstill waterですか？”
(Is this still water?)



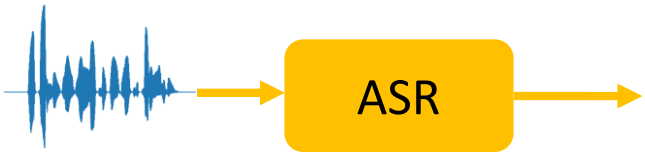
Previous Works

Several works have constructed CS ASR.

- Mandarin-English CS with phone merging and language identification [Vu et al., 2012]
- Frisian-Dutch CS with bilingual deep neural networks [Yilmaz et al., 2012]

⇒ Common aim: merely for transcribing CS-speech into CS-text

CS-speech



CS-text

“父の手術には there is a high degree of risk.”

(there is a high degree of risk in my father's operation)

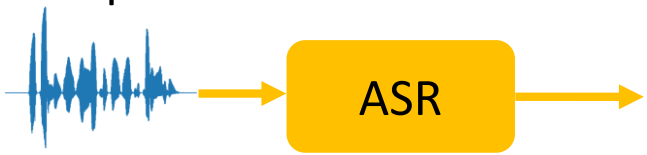
Monolingual speakers

Proposed Approaches

Goal

Support monolingual speakers trying to understand CS speakers

CS-speech



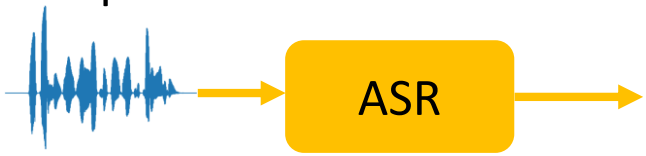
CS-text

“父の手術には there is a high degree of risk.”

(there is a high degree of risk in my father's operation)

Monolingual speakers

CS-speech



Monolingual-text

“there is a high degree of risk in my father's operation.”

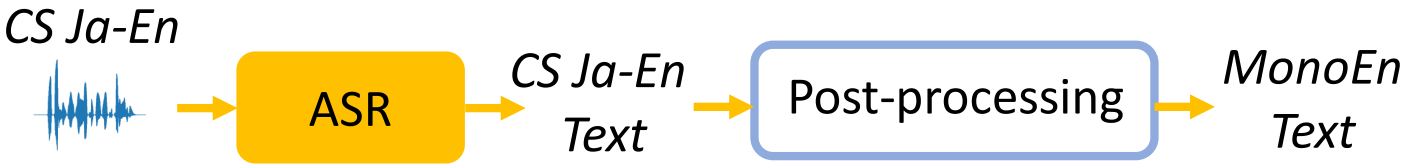
Monolingual speakers

Approaches

1. Cascade approaches

1-1. CS2CS ASR + Mono-recovery BERT

1-2. CS2CS ASR + CS2Mono NMT



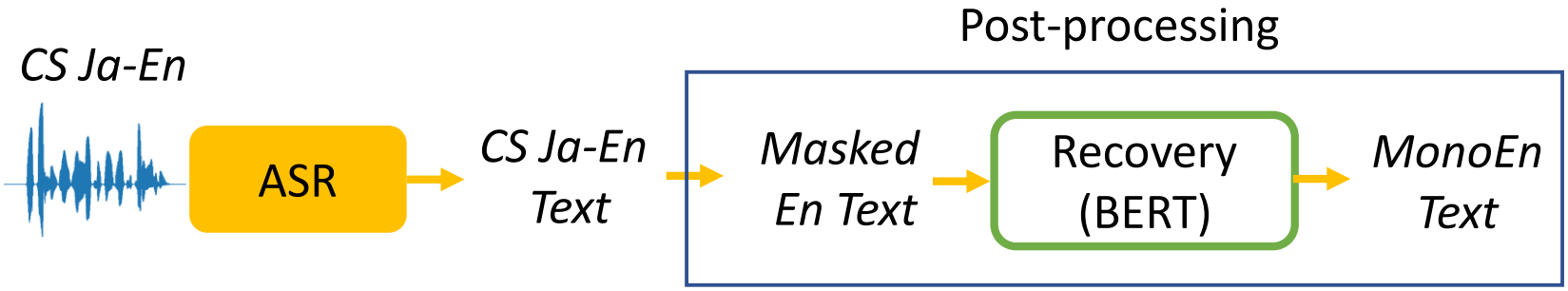
2. Direct approaches

2-1. CS2Mono ASR with single-task learning

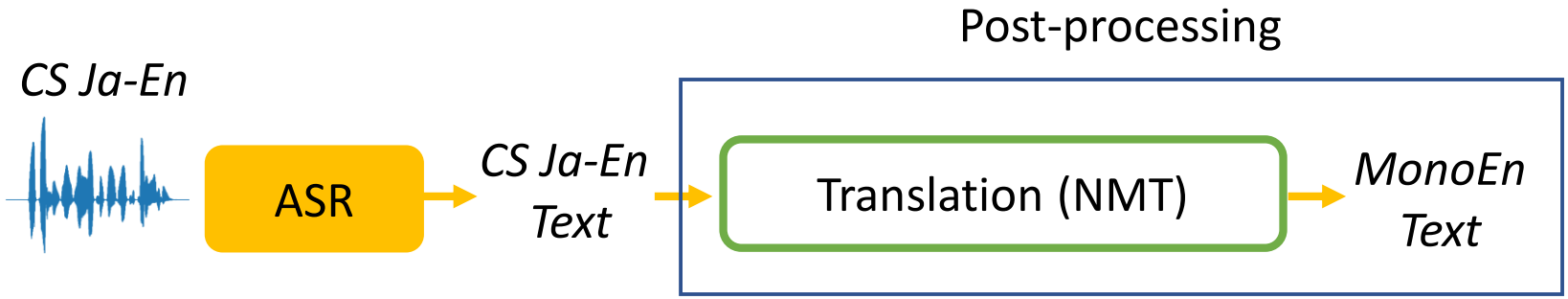
2-2. CS2Mono ASR with multi-task learning

1. Cascade Approaches

1-1. Cascade CS2CS ASR + Mono-recovery BERT



1-2. Cascade CS2CS ASR + CS2Mono NMT*

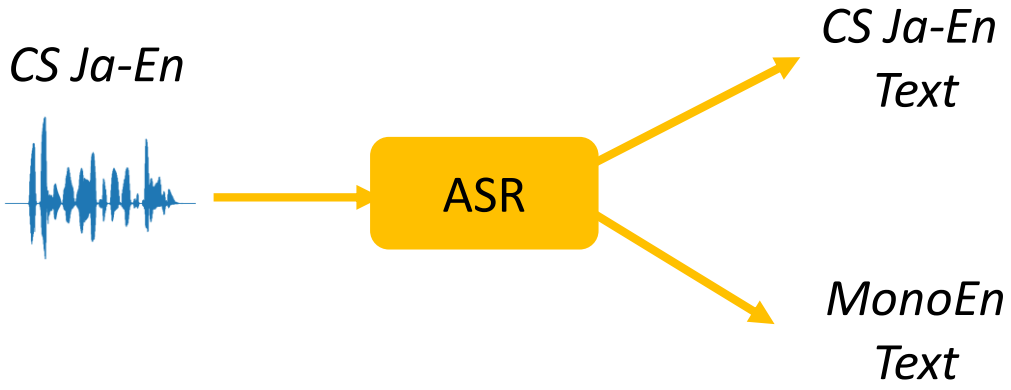


*Neural Machine Translation (NMT)

2. Direct Approaches

2-1. Direct CS2Mono ASR with single-task learning

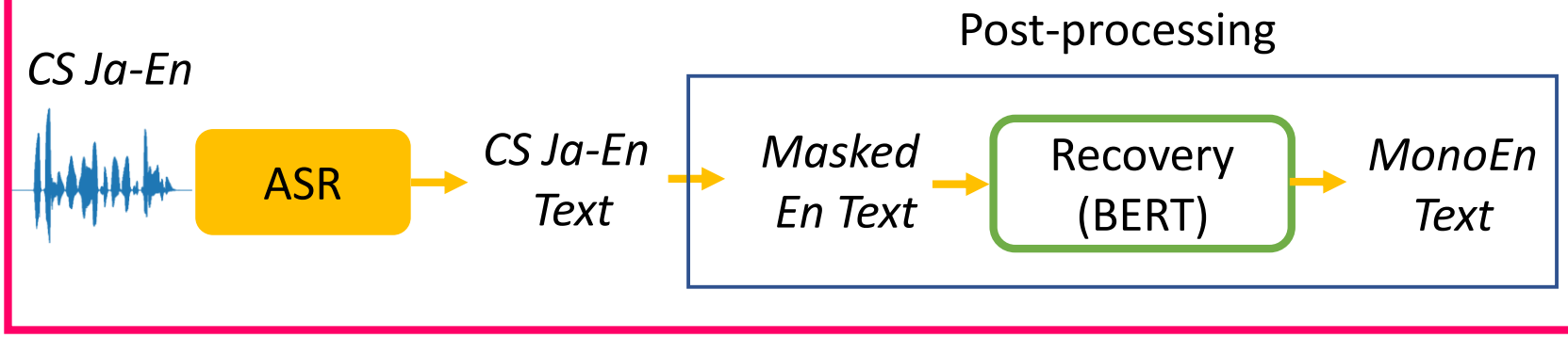
2-2. Direct CS2Mono ASR with multi-task learning



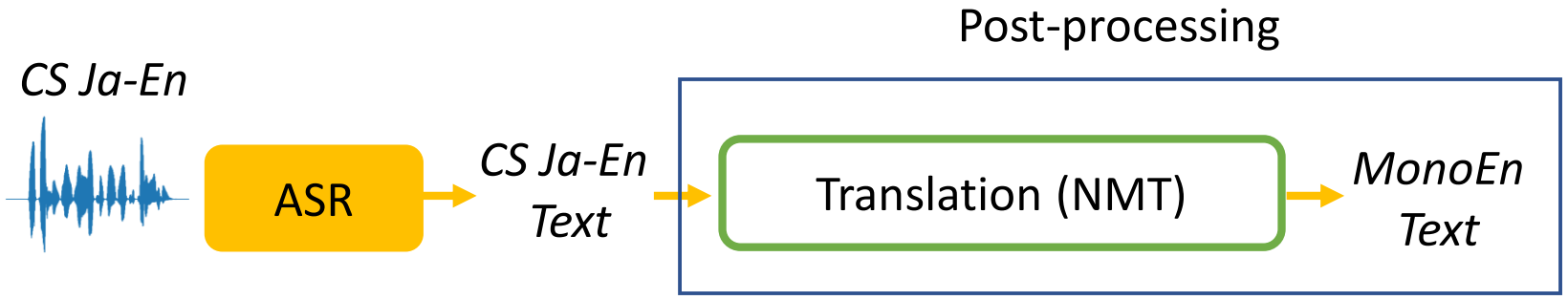
⇒ I will describe these proposed models one by one

1. Cascade Approaches

1-1. Cascade CS2CS ASR + Mono-recovery BERT



1-2. Cascade CS2CS ASR + CS2Mono NMT*



*Neural Machine Translation (NMT)

1-1. Mono-Recovery BERT

BERT: bidirectional language model [Devlin et al., 2019]

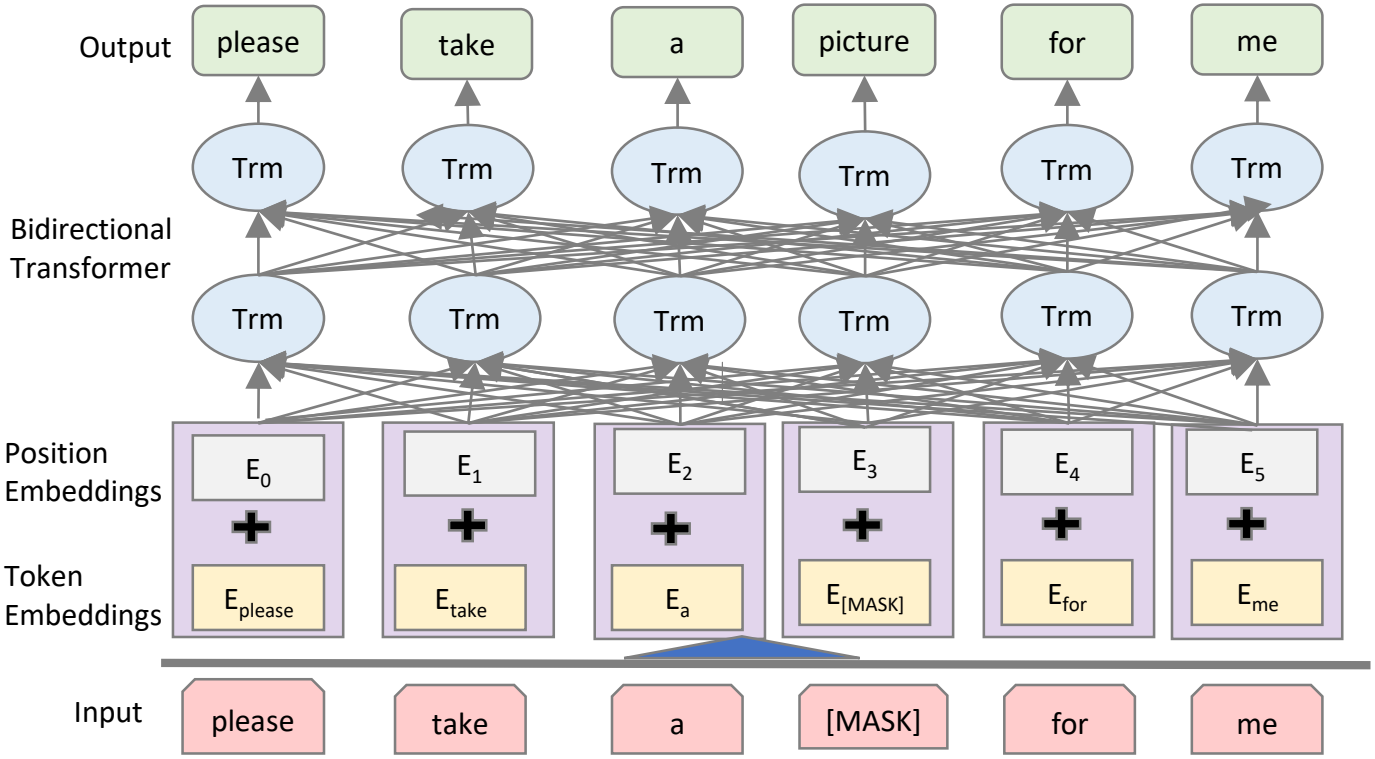
Mono-recovery BERT

- ☐ Masks the 2nd language
- ☐ Recovers complete sentence of the 1st language

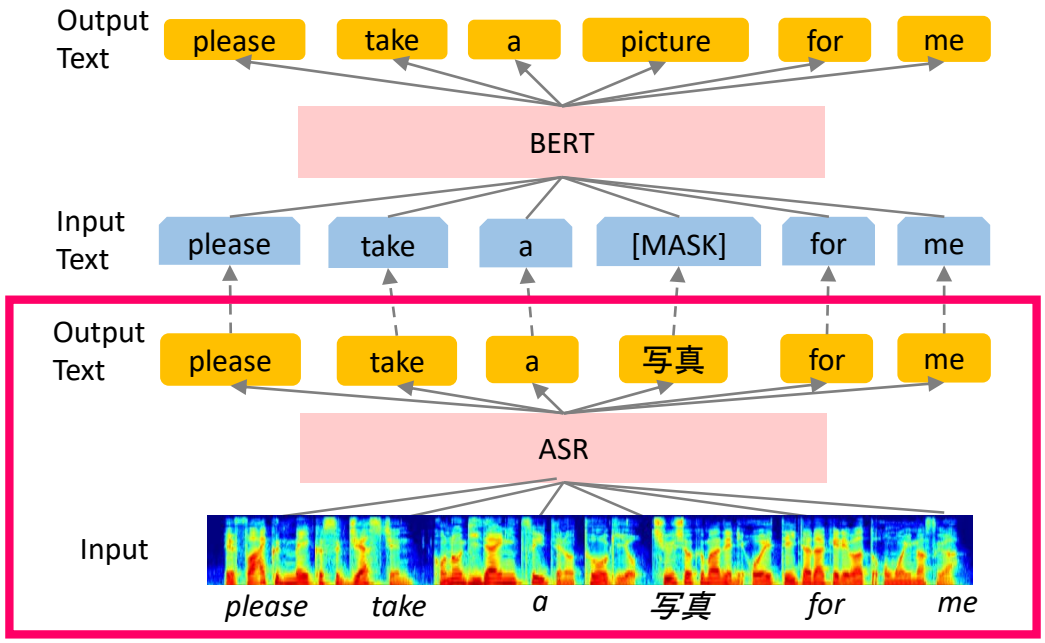
Source	あのね， Charles may be a bad husband, but He's 心の温かい人なのよ. *(you know, Charles may be a bad husband, but He's a warm-heated person.)*
Mask	[MASK] [MASK] [MASK], Charles may be a bad husband, but He's [MASK] [MASK] [MASK].
Label	you know [PAD], Charles may be a bad husband, but He's a warm-heated person.
Target	you know, Charles may be a bad husband, but He's a warm-heated person.

1-1. Mono-Recovery BERT Architecture

- Based on the BERT$_{Base}$ model
  - Use multi-layer bidirectional Transformer [Vaswani et al., 2017]



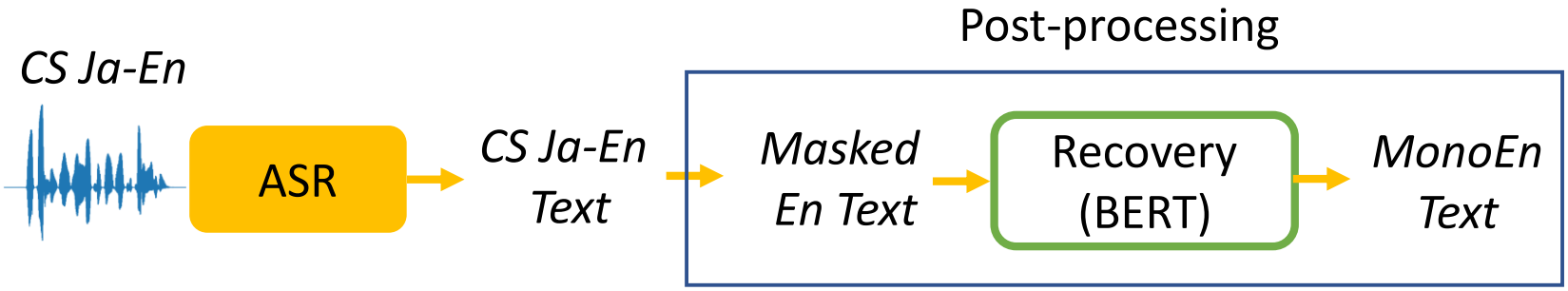
Overview of Cascade BERT



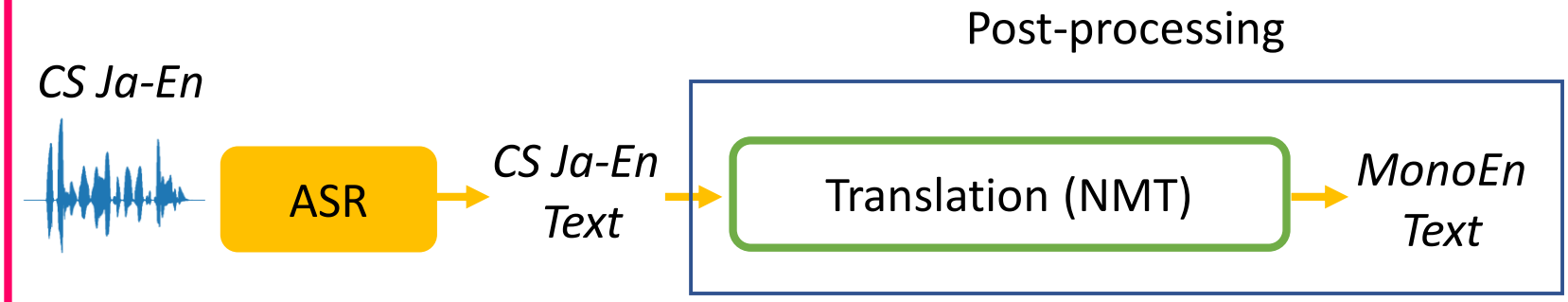
Given CS speech, we performs an ASR and produces CS text.
Then, we utilizes BERT to recover the monolingual text.

1. Cascade Approaches

1-1. Cascade CS2CS ASR + Mono-recovery BERT



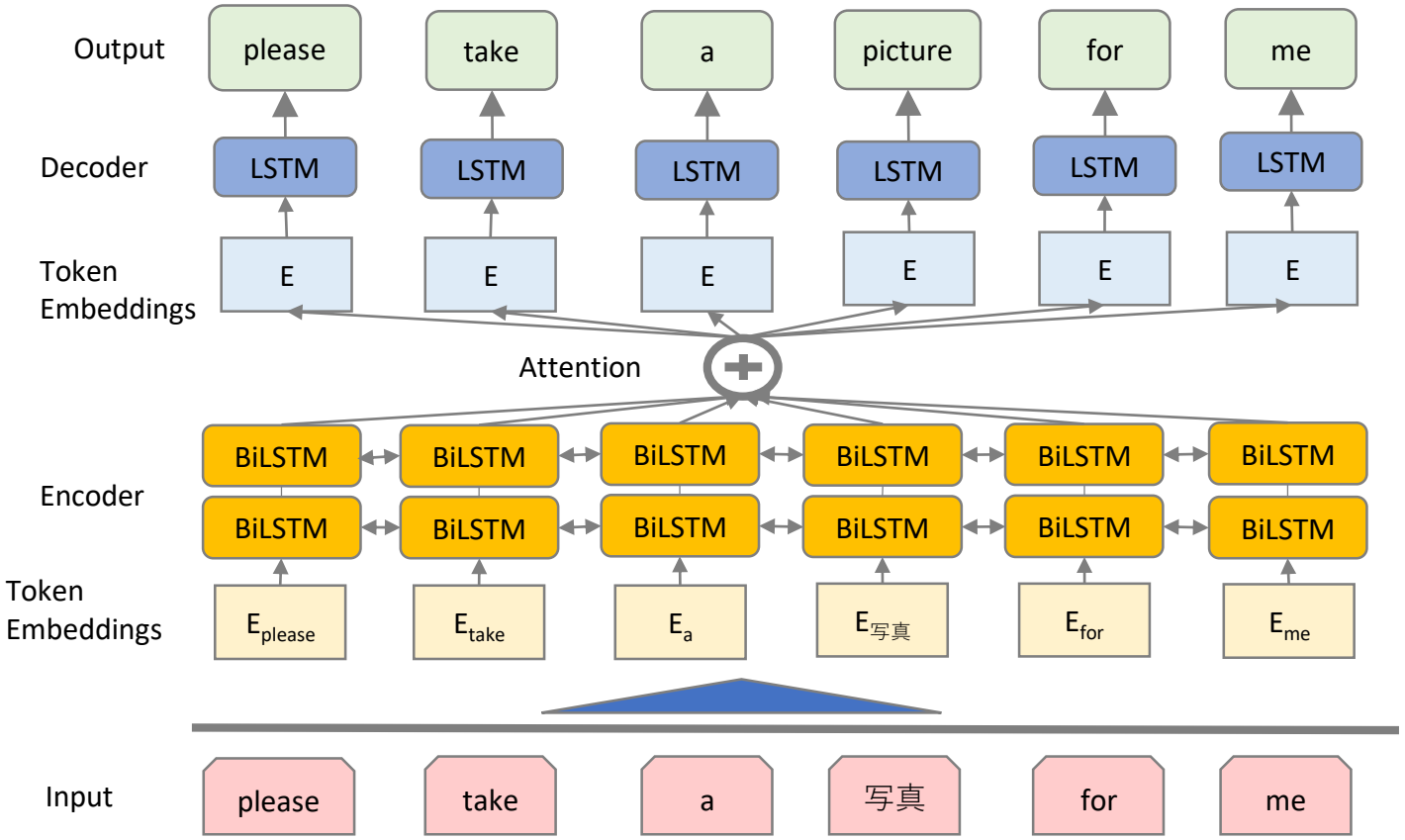
1-2. Cascade CS2CS ASR + CS2Mono NMT*



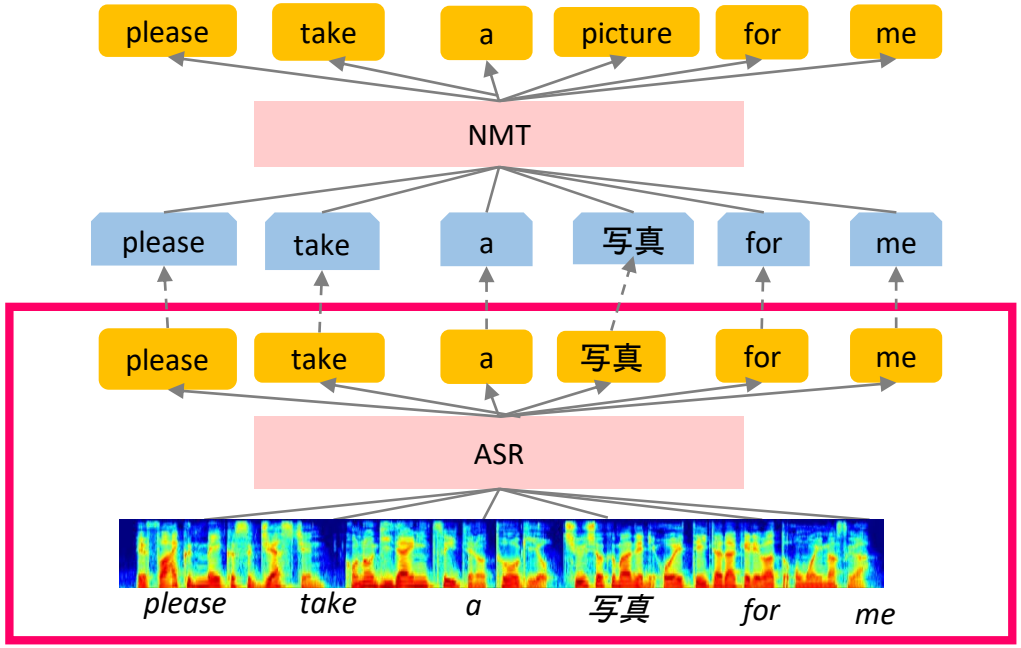
*Neural Machine Translation (NMT)

1-2. NMT Model

- Sequence-to-sequence NMT with attention [Bahdanau et al., 2015]



Overview of Cascade NMT

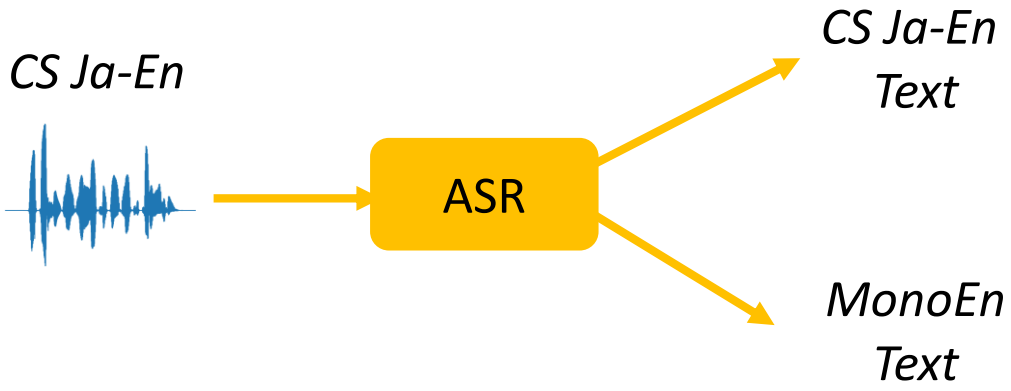


Given CS speech, we performs an ASR and produces CS text.
Then, we utilizes NMT to translate from CS to monolingual text.

2. Direct Approaches

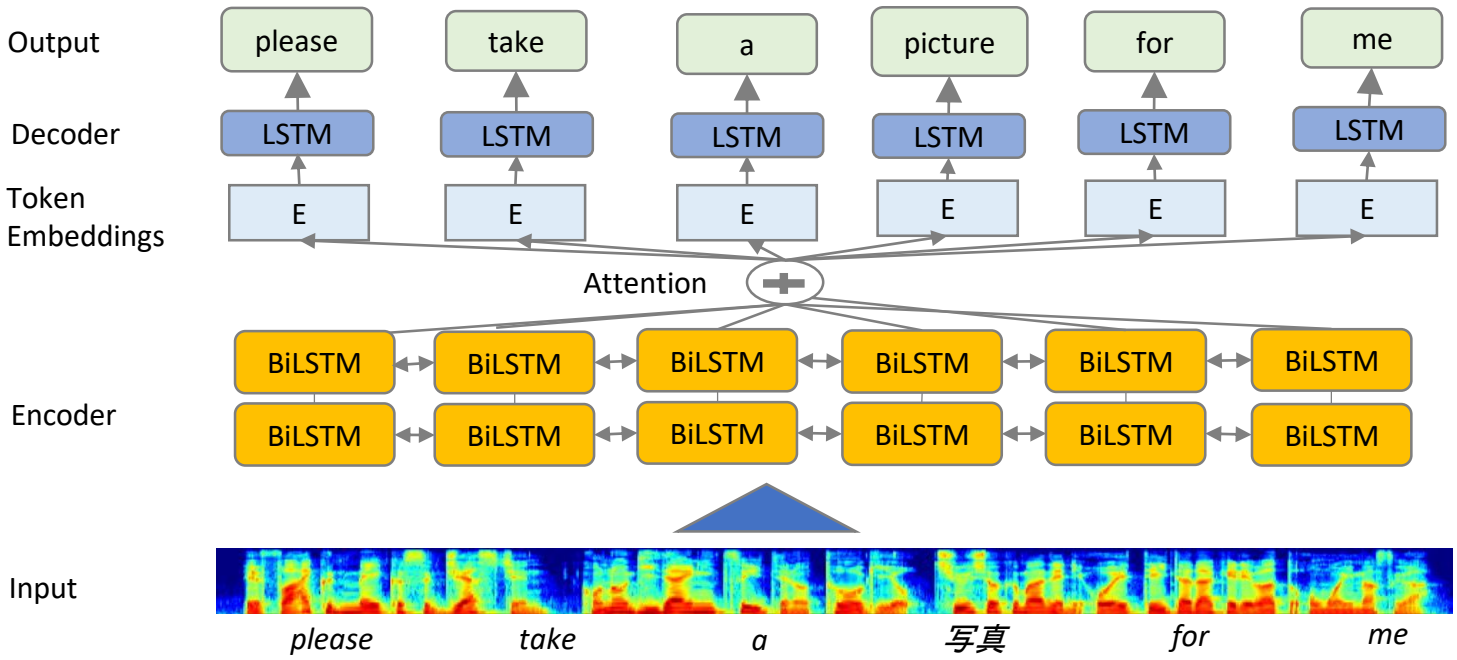
2-1. Direct CS2Mono ASR with single-task learning

2-2. Direct CS2Mono ASR with multi-task learning



2-1. Direct Single-task Learning

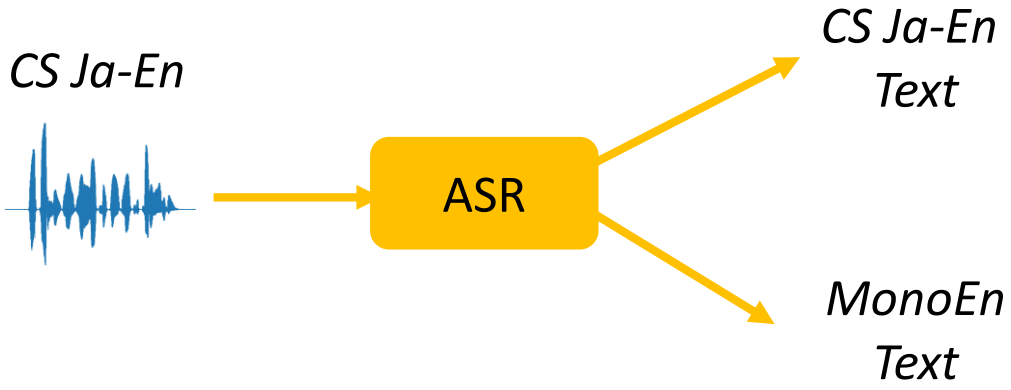
- Sequence-to-sequence ASR with attention [Chan et al., 2016] [Tjandra et al., 2017]



2. Direct Approaches

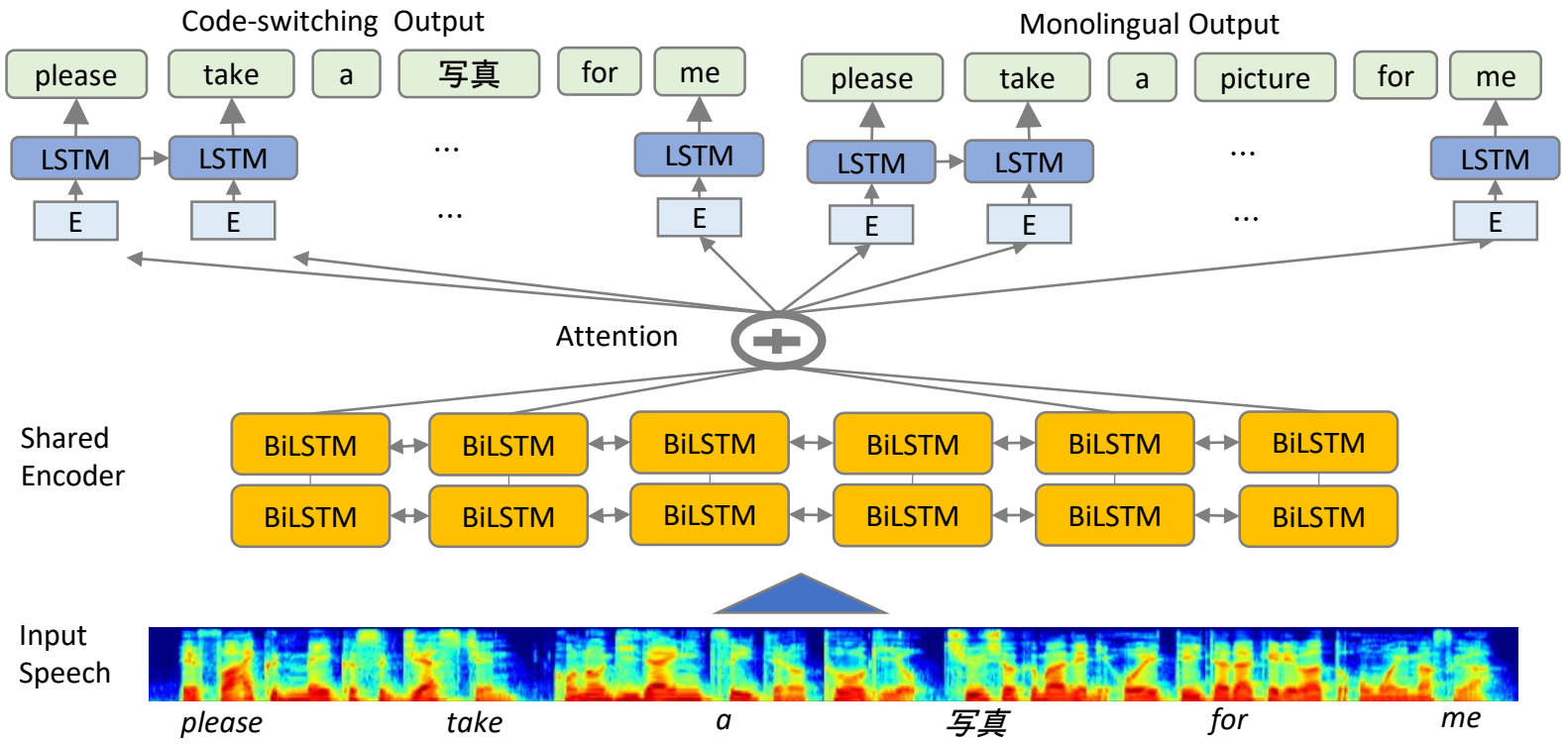
2-1. Direct CS2Mono ASR with single-task learning

2-2. Direct CS2Mono ASR with multi-task learning



2-2. Direct Multi-task Learning

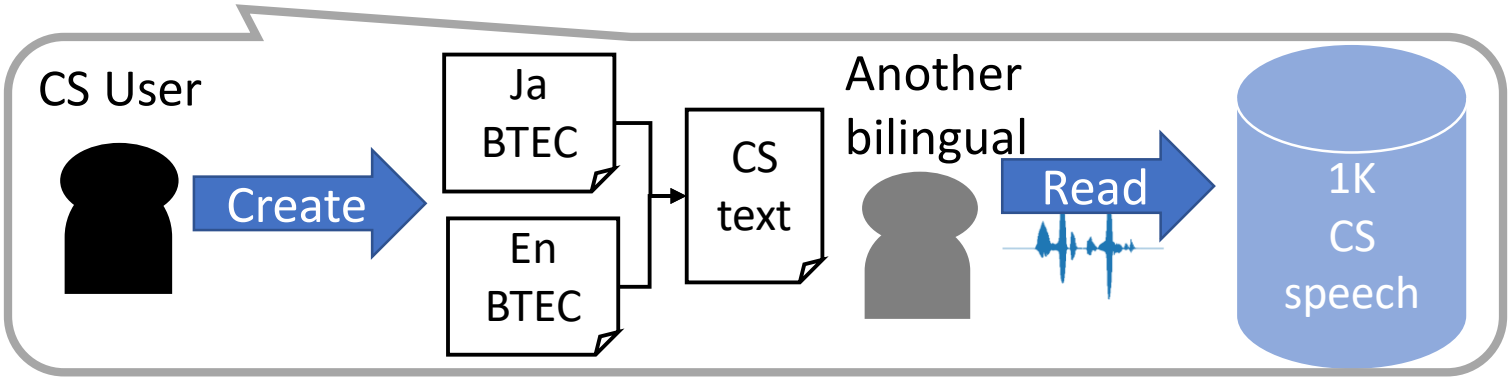
- Multi-task learning improves learning efficiency with a shared encoder.



Experiments

Datasets

Artificial speech	CS text created by translating part of BTEC
	Google TTS speech
Natural speech	CS text created by a CS user
	Bilingual reading speech



- All artificial CS : Train 50K, Test 500
- Mix natural CS : Train 50K, Test 500

All texts are segmented with BERT wordpiece tokenizer.
Wordpiece: subword units proposed to effectively deal with rare words

Cascade Text-to-Text Results

Pre-experiment: check the difference between with ASR error and without

Wordpiece error rate* (%) of phrase-level CS on BERT model

	Artificial CS Test	Mix Natural CS Test
All Artificial CS	9.75	14.96
Mix Natural CS	10.37	13.21

Wordpiece error rate (%) of phrase-level CS on NMT model

	Artificial CS Test	Mix Natural CS Test
All Artificial CS	7.56	16.34
Mix Natural CS	6.34	19.16

$$*Wordpiece\ Error\ Rate[\%] = \frac{Total\ Wordpiece\ Errors}{Number\ of\ Wordpieces\ in\ Reference} \times 100$$

- NMT model translated all artificial CS test utterances well, but the BERT model translated the mixed natural CS test utterances better.
- Natural CS has more complex Japanese phrases, so the NMT task became more difficult.

Cascade Speech-to-Text Results

Wordpiece error rate (%) of phrase-level CS on cascade of ASR+BERT

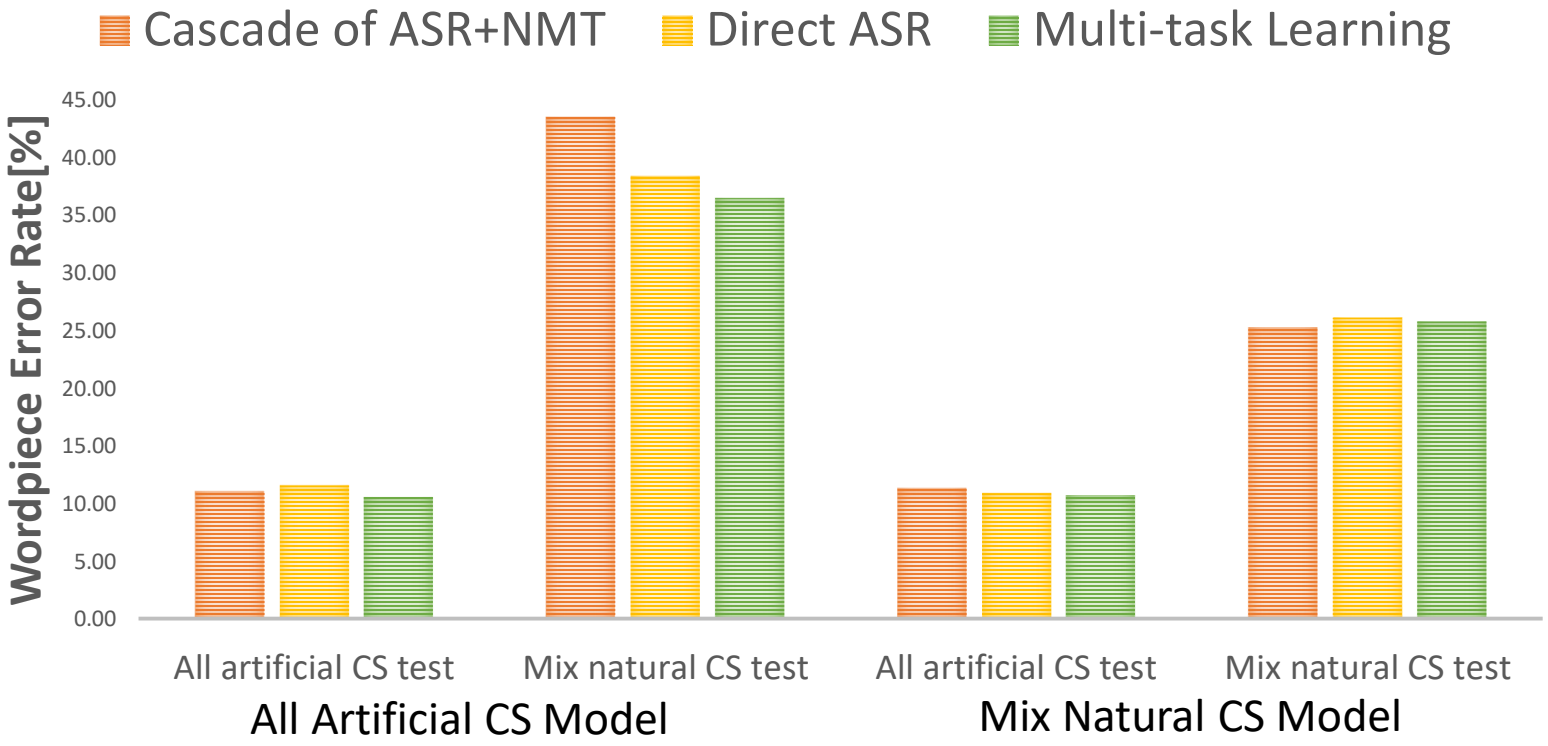
	Artificial CS Test	Mix Natural CS Test
All Artificial CS	17.80	47.98
Mix Natural CS	18.62	29.22

Wordpiece error rate (%) of phrase-level CS on cascade of ASR+NMT

	Artificial CS Test	Mix Natural CS Test
All Artificial CS	11.07	43.52
Mix Natural CS	11.32	25.29

- However, in speech-to-text, NMT outperformed BERT.
- It seems a more difficult task for BERT because the ASR error increased the number of [MASK] tokens.
- Therefore, we compare the cascade of ASR+NMT with other models.

Cascade and Direct Results



- Direct approaches seem to be better than the cascade model.
  - They can learn speech information directly.
- Multi-task learning tends to have the best performance.
  - It can improve learning efficiency and accuracy.

Conclusion

- We compared four ways to translate CS speech:

1. Cascade CS2CS ASR + mono-recovery BERT
2. Cascade CS2CS ASR + CS2Mono NMT
3. Direct CS2Mono ASR with single-task learning
4. Direct CS2Mono ASR with multi-task learning

- The results reveal that
  - ASR error makes the task difficult for mono-recovery BERT.
  - Direct approaches seem to be better than the cascade model.
  - Multi-task learning tends to have the best performance.
- In the future, we will further investigate how natural the translation results are by conducting an evaluation by bilingual people.
- As this paper is targeted only to proficiency-driven CS, we will handle deficiency-driven CS in the future.

References

- Jeff McSwan, “The architecture of the bilingual language faculty: Evidence from intrasentential code-switching,” Bilingualism: Language and Cognition, vol. 3, no. 1, pp. 37–54, 2000.
- Maria Lourdes S Bautista, “Tagalog-english code switching as a mode of discourse,” in Proc. of Asia Pacific Education Review, 2004, vol. 5, pp. 226–233.
- Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Eng-Siong Chng, Tanja Schultz, and Haizhou Li, “A first speech recognition system for Mandarin-English code-switch conversational speech,” in Proc. of ICASSP, Kyoto, Japan, 2012, pp.4889–4892.
- Emre Yilmaz, Henkvan den Heuvel, and David van Leeuwen, “Investigating bilingual deep neural networks for automatic recognition of codeswitching Frisian speech,” in Proc. of SLTU, Yogyakarta, Indonesia, 2016, vol. 81, pp. 159 – 166.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in Proc. of NAACL-HLT, 2019, pp. 4171–4186.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in Proc. of NIPS, 2017, pp. 5998–6008.
- Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Proceedings of ICLR, 2015.
- William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 4960–4964.
- Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura, “Listening while speaking: Speech chain by deep learning,” CoRR, vol. abs/1707.04879, 2017. [Online]. Available: http://arxiv.org/abs/1707.04879
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. “Google’s neural machine translation system: Bridging the gap between human and machine translation,” CoRR, vol. abs/1609.08144, 2016.