ICT's System for AutoSimTrans 2021: Robust Char-Level Simultaneous Translation

Shaolei Zhang [1,2], **Yang Feng** [1,2]*
[1]Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] University of Chinese Academy of Sciences, Beijing, China
{zhangshaolei20z, fengyang}@ ict.ac.cn

中国科学院大学
University of Chinese Academy of Sciences

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Contents

Motivation

- ▷ **Pipeline of simultaneous interpretation**
  - Automatic Speech Recognition (ASR) →
    simultaneous translation (ST) →
    Text-to-Speech Synthesis (TTS)
- ▷ **Input of simultaneous translation:**
  - Inaccurate, unsegmented.
  - Spoken language domain.
- ▷ **Robustness** **and** **Domain adaptability**

Streaming Transcript	Translation
大	
大家	
大家好	Hello everyone!
欢	
欢迎	
欢迎大	
欢迎大家	
欢迎大家来	Welcome
欢迎大家来到	everyone
欢迎大家来到这	come
欢迎大家来到这里	here.

Motivation

▷ **For robustness**

- ASR result (streaming transcription): incremental, unsegmented.
- Subword-level segmentation result of the streaming transcription is unstable.
  - Existing method: remove the last to prevent it from being incomplete.

Streaming Transcription	Tokenization of Streaming Transcription Input		
	Standard Wait-2	Standard Wait-2 Remove Last Token	Char-Level Wait-2 (Ours)
他是研究生物的		No input	
他是研究生物的	他/是/ Unstable input	他/	他/是/
他是研究生物的	他/是/研/	他/是/ ▷ Delay	他/是/研/
他是研究生物的	他/是/研究/ ✘	他/是/ ▷ Delay	他/是/研/究/
他是研究生物的	他/是/研究生/ ✘	他/是/ ▷ Delay	他/是/研/究/生/
他是研究生物的	他/是/研究/生物/ ✘	他/是/研究/	他/是/研/究/生/物/
他是研究生物的	他/是/研究/生物/的/	他/是/研究/生物/的/	他/是/研/究/生/物/的/

Motivation

- **For domain adaptability**
  - General domain the spoken language domain are quite different:
    - Word order
    - Punctuation
    - Modal particles
    - …

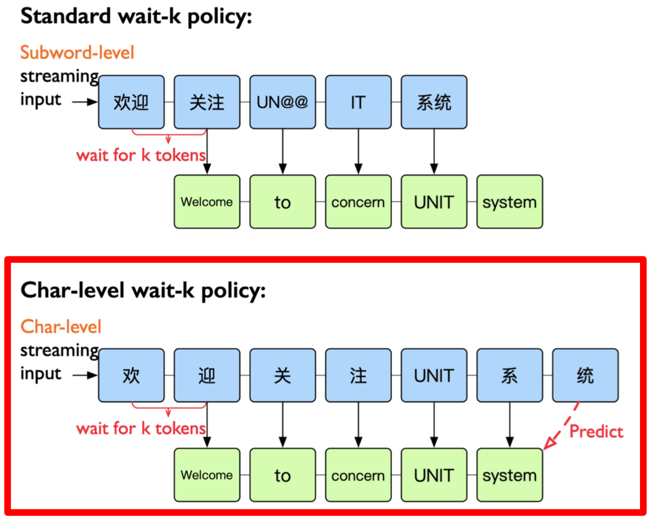
- **Our system**
  - Robust:
    - Propose the Char-Level Wait-k Policy
  - Domain adaptation:
    - Apply data augmentation on spoken language domain.
    - Combine two training methods to enhance the predictive ability.

Contents

Method

▷ **Char-Level Wait-k Policy**

- Source: character sequence after char-level tokenization.
- Target: subword sequence after subword-level segmentation and BPE.
- Read / Write policy: waiting for k source characters first, and then reading and writing alternately.



Input Sentence		欢迎来到UNIT系统的第12期高级课程。
Output Sentence		welcome to the 12th advanced course on UNIT system .
S.	subword-level MT	欢迎/来到/UN@@/IT/系统/的/第/12@@/期/高级/课程/。
	character-level MT	欢/迎/来/到/U/N/I/T/系/统/的/第/1/2/期/高/级/课/程/。
Ours	char-level tokenization	欢/迎/来/到/UNIT/系/统/的/第/12/期/高/级/课/程/。
T.	subword-level MT	welcome/to/the/12@@/th/advanced/course/on/UNIT/system/.

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Method

- **Why char-level simultaneous translation?**
  - **More robust**
    - avoid unstable prefixes caused by subword segmentation.
  - **More fine-grained latency**
    - if one character is enough to express the meaning of a entire word, the ST system does not have to wait for the complete word.
  - **Translation quality will not be affected too much**
    - only performs char-level tokenization on the source, and the target retains subword-level tokenization.

Method

▷ Domain Adaptation

- Depunctuation
  - Source: delete the ending punctuation.
  - Target: unchanged.

Original	各位开发者、各位朋友们，大家下午好！
Depunctuation	各位开发者、各位朋友们，大家下午好

- Data Augmentation
  - For spoken language domain corpus.
  - Source: we perform 5 data augmentation operations.
  - Target: unchanged.

Original	1957年我到北京上大学
Add Comma	1957年，我到北京上大学
Add Tone character	1957年我啊到北京上大学
Copy Character	1957年我到北北京上大学
Replace Homophone	1957年我到北经上大学
Delete Character	1957年我到北京上大学

中国科学院计算技术研究所 INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

Method

- **Training Methods**
  - Pre-training : general domain MT corpus
    - Multi-path training (Elbayad et al., 2020)
    - Future-guided training (Zhang et al., 2020b)
  - Fine-tuning : spoken language domain corpus
    - Original training: fix k and use the original prefix-to-prefix framework for training, and train different models for different k.

Maha Elbayad, Laurent Besacier, and Jakob Verbeek. 2020. Efficient wait-k models for simultaneous machine translation.

Shaolei Zhang, Yang Feng, and Liangyou Li. 2020b. Future-guided incremental transformer for simultaneous translation.

Contents

Experiments

▷ Datasets

- CWMT19 Chinese → English: for pre-training.
- Transcription: for fine-tuning.
- Dev. Set: for evaluation.

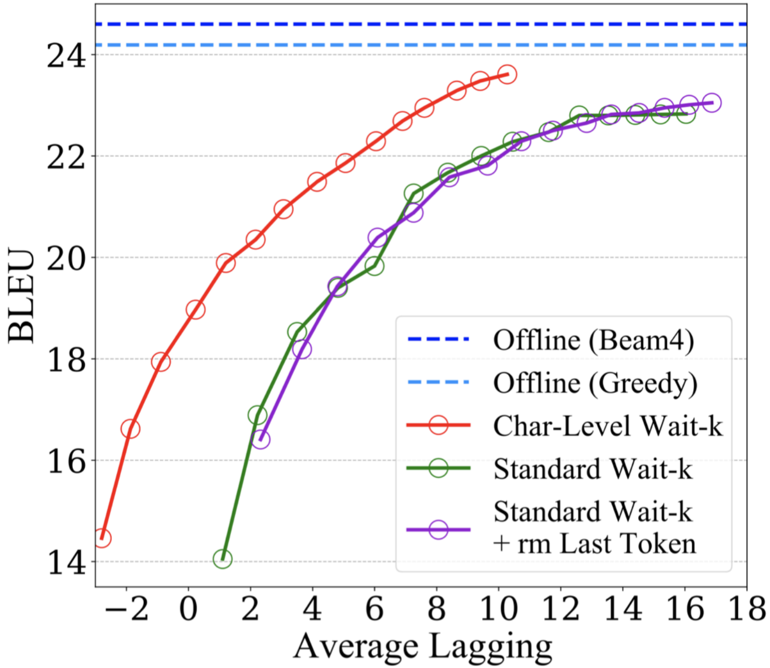
Datasets	Domain	#Sentence Pairs
CWMT19	General	9,023,708
Transcription	Spoken	37,901
Dev. Set	Spoken	956

▷ System setting

- Offline: full-sentence MT based on Transformer.
- Standard Wait-k: standard subword-level waitk policy.
- Standard Wait-k + rm Last Token: In the inference time, the last token after the word segmentation is remove to prevent it from being incomplete.
- Char-Level Wait-k: our proposed method.

Experiments

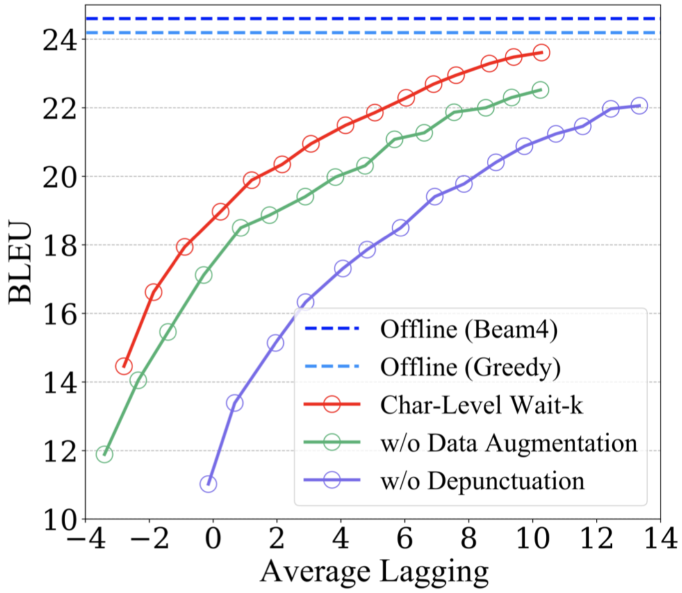
- **Main Result**
  - Char-Level Wait-k improves about 6 BLEU at low latency (AL=1.10).
  - More stable and robust.

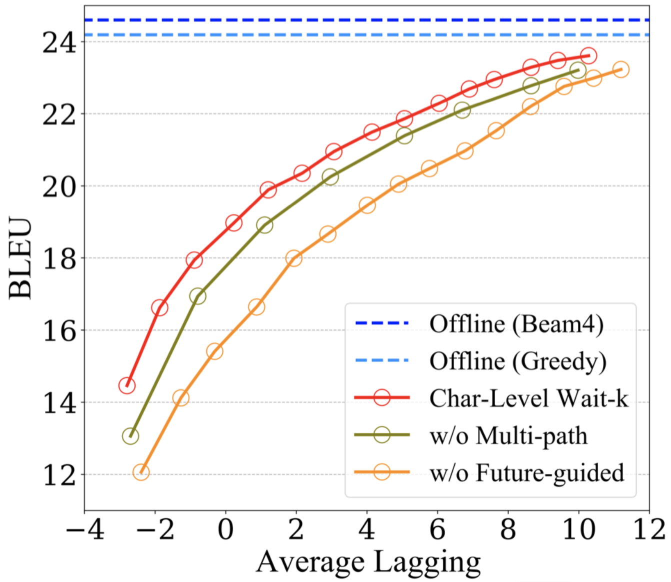


Experiments

▷ **Ablation Study**

- Data processing: 'Depunctuation' and 'Data Augmentation'
- Training methods: 'Future-guided' and 'Multi-path'





Contents

Conclusion

- The proposed char-level wait-k policy is more robust.
- Data processing and two training methods improve the spoken language domain adaptability.

- For some language pairs with a large length ratio between the source (char) and the target (bpe), we can read multiple characters at each step to deal with the long char-level source. We put this into our future work.

Thanks!

Contact me with:

zhangshaolei20z@ict.ac.cn