SIMQA: Detecting Simultaneous MT Errors through Word-by-Word Question Answering

HyoJung Han, Marine Carpuat, Jordan Boyd-Graber
University of Maryland, College Park in Department of Computer Science

Summary

- In current eval of Simultaneous Machine Translation (**SimulMT**, **trade-off** between **quality** and **latency** does not fully reflect timely adequacy.
- We introduce a Cross-lingual word-by- word question answering task, **SimQA** to quantify the **timely adequacy** of SimulMT more **directly**.
- We construct the Cross-lingual Quizbowl test set, **XQB** by translating Polish and Spanish questions and answers into English.
- Our SimQA results complement intrinsic QA and MT metrics by jointly accounting for **timeliness** and **translation quality**.
- We suggest that SimQA can diagnose critical SimulMT errors on the fly.

Motivations

Evaluation of SimulMT is Hard!

Simultaneous Machine Translation starts translating prefix of source text before the entire source text is available.

Step	Input Source	Decision	Target Output
t	... s_i	Read	... t_j
t+1	... s_i s_{i+1}	Write	... t_j
t+2	... s_i s_{i+1}	...	... t_j t_{j+1}

Current prevailing Method of SimulMT Evaluation : Quality + Latency

Quality
- Full-input based standard metrics
- Fail to capture salient meaning errors
- Not suited for SimulMT (dropping or simplifying can be done)
- Quality as "perfection" rather than "**fitness for purpose**"

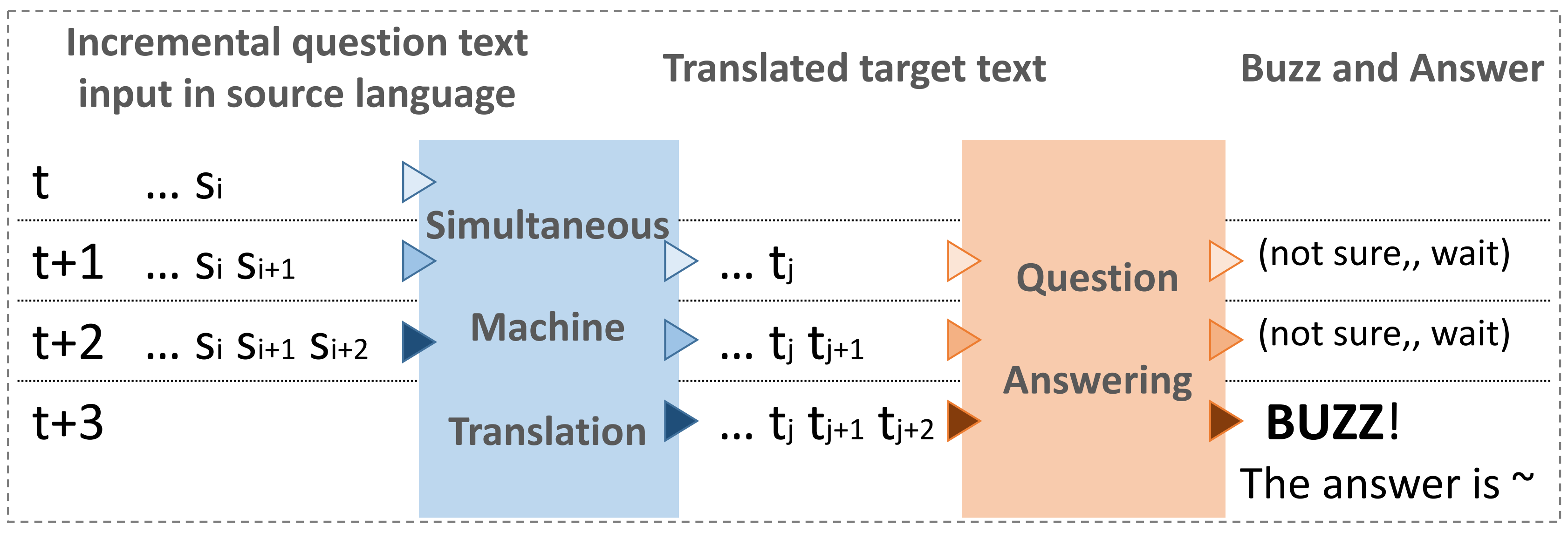
Latency
- Still hard to know "What degree SimulMT translation are useful for practical purpose?"

QuizBowl System as a **proxy** task for eval of SimulMT, since also deals with incremental inputs and with sequential decision making. Based on produced guesses, Buzzer decide whether to buzz or not. The goal is to buzz with correct answer as soon as possible.

Step	Input Q text	Guesses (top N)	Buzz?
t	... t_j	$\{A^t_1, \dots, A^t_N\}$	no
t+1	... t_j t_{j+1}	$\{A^{t+1}_1, \dots, A^{t+1}_N\}$	no
t+2	... t_j t_{j+1} t_{j+2}	$\{A^{t+2}_1, \dots, A^{t+2}_N\}$	Yes!

SimQA: Cross-lingual Word-by-Word QA task

Task driven evaluation of SimulMT based on **word-by-word QA**
Directly measures the quality of **timely adequacy**



Experiment Settings

International Academic Competitions

XQB
- **Cross-lingual Quizbowl Dataset**
- Collection of **Polish** (#512) and **Spanish** (#148) (feat. IAC)
- English Reference by Human for Question and Answer

Model
- SimulMT : Wait-K[2] model {k=3,6,9,12}, Trained on WMT
- QuizBowl QA: Guesser (GRU, Bert, Elastic Search), Buzzer (LSTM)

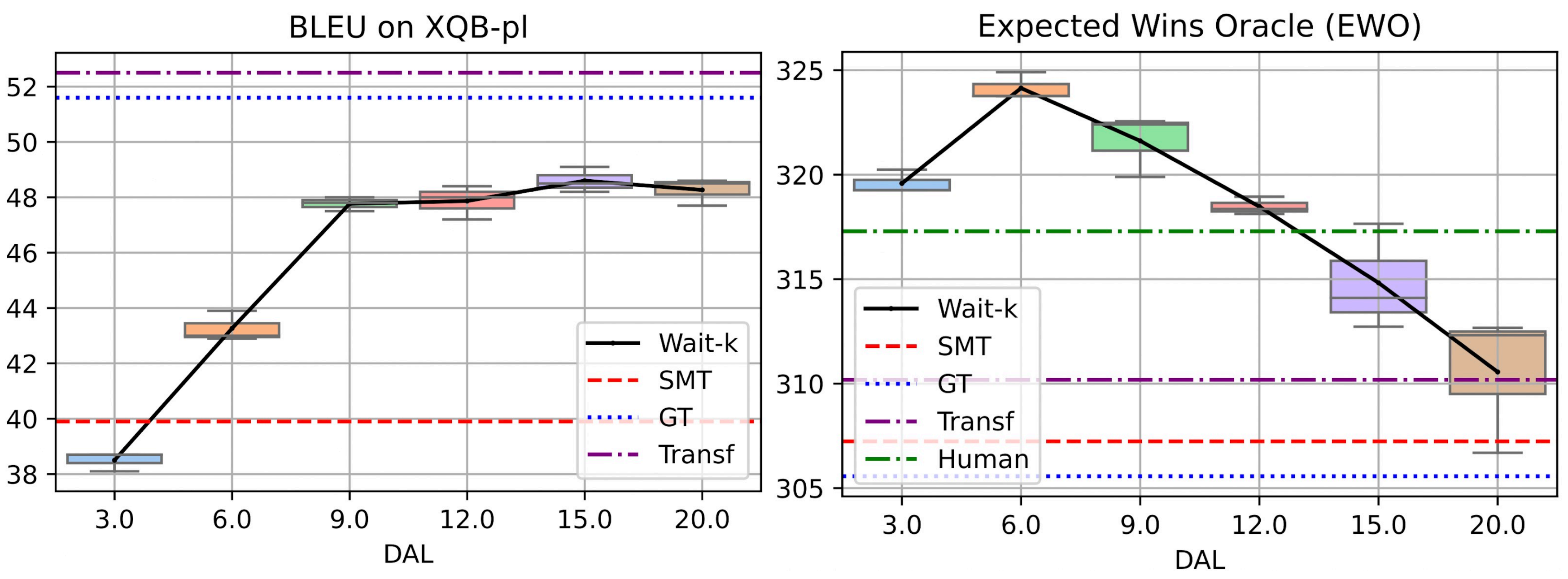
Metric
- QA : Expected Win (EW), EW with Optimal Buzzer (EWO)
- MT : BLEU[3], COMET[4], BertScore[5], …
- Latency : Differentiable Average Lagging (DAL)[6]

Results

SimQA results vs MT metrics

Traditional metrics of MT quality all increase monotonically. (BLEU)

By **jointly** accounting for timeliness and translation quality, SimQA evaluation reveals different trends and peaks at Wait-6. (EWO)



Analysis

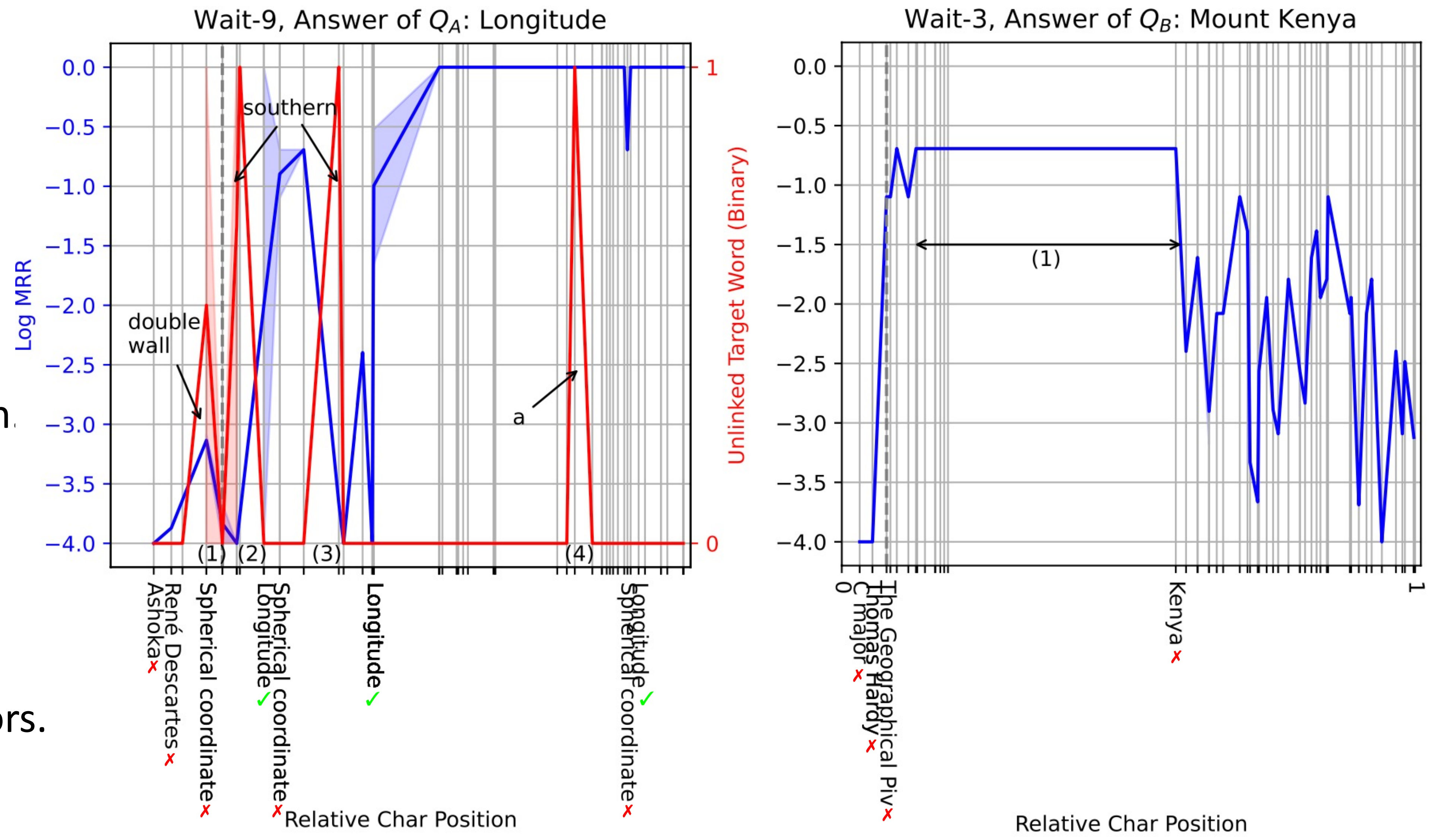
Stepwise Visualization of SimulMT Errors

We also analyzed the behavior of the SIMQA system step-by-step, and see how it can reveal critical errors like hallucinations or under-translation.

1. Hallucination : The position of significant drop in log Mean Reciprocal Rank (MRR) correlates with hallucinated word position.

2. Under-Translation : Flat log MRR over long range of relative position correlates with under-translation errors.

QA model provide useful signals to pinpoint critical SimulMT errors.



Reference

[1]Quizbowl: The Case for Incremental Question Answering (Rodriguez et al., Arxiv 2019)
[2]STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework (Ma et al., ACL 2019)
[3]Bleu: a Method for Automatic Evaluation of Machine Translation (Papineni et al., ACL 2002)
[4]COMET: A Neural Framework for MT Evaluation (Rei et al., EMNLP 2020)
[5]BERTScore: Evaluating Text Generation with BERT (Zhang et al., ICLR 2020)
[6]Monotonic Infinite Lookback Attention for Simultaneous Machine Translation (Arivazhagan et al., ACL 2019)