**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.

<table>
<thead>
<tr>
<th>Step</th>
<th>Input Source</th>
<th>Decision</th>
<th>Target Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>... $S_t$</td>
<td>Read</td>
<td>$t_1$</td>
</tr>
<tr>
<td>t+1</td>
<td>... $S_{t+1}$</td>
<td>Write</td>
<td>... $t_1$</td>
</tr>
<tr>
<td>t+2</td>
<td>... $S_{t+1}$</td>
<td>...</td>
<td>... $t_1$ $t_2$</td>
</tr>
</tbody>
</table>

Current prevailing Method of SimulMT Evaluation: Quality + Latency

- 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”
- 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.

<table>
<thead>
<tr>
<th>Step</th>
<th>Input Q text</th>
<th>Guesses (top N)</th>
<th>Buzz?</th>
</tr>
</thead>
<tbody>
<tr>
<td>t</td>
<td>... $t_1$</td>
<td>$A_{1,...,t}$</td>
<td>no</td>
</tr>
<tr>
<td>t+1</td>
<td>... $t_1$ $t_2$</td>
<td>$A_{1,...,t_2}$</td>
<td>no</td>
</tr>
<tr>
<td>t+2</td>
<td>... $t_1$ $t_2$ $t_3$</td>
<td>$A_{1,...,t_3}$</td>
<td>yes</td>
</tr>
</tbody>
</table>

**Analyses**

**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.

**Experiment Settings**

- **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)
- QA : Expected Win (EW), EW with Optimal Buzzer (EWO)

**Metric**

- 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**

Wait-9, Answer of Q$_9$, Longitude

Wait-3, Answer of Q$_3$, Mount Kenya

**Reference**