HyoJung Han, Marine Carpuat, and **Jordan Boyd-Graber**. **SimQA: Detecting Simultaneous MT Errors through Word-by-Word Question Answering**. *Empirical Methods in Natural Language Processing*, 2022, 9 pages.


```
    @inproceedings{Han:Carpuat:Boyd-Graber-2022,
Title = {SimQA: Detecting Simultaneous MT Errors through Word-by-Word Question Answering},
Author = {HyoJung Han and Marine Carpuat and Jordan Boyd-Graber},
Booktitle = {Empirical Methods in Natural Language Processing},
Year = {2022},
Location = {Abu Dhabi},
Url = {http://cs.umd.edu/~jbg//docs/2022_emnlp_simqa.pdf},
}
```


Accessible Abstract: Simultaneous interpretation (where a translation happens word by word before the source sentence is finished) is difficult to evaluate. We created a new evaluation framework based on the following scenario: imagine that you're thrown into a trivia gameshow where you don't know the language. Specifically, it's a game format where you interrupt the question word by word as soon as possible. Our hypothesis is that a monolingual player (who doesn't speak the source language) will be able to do better in the game with a better simultaneous translation system. In this 2022 EMNLP publication, we show that this evaluation is not only cheaper (you just need to translate the answer) but can also detect hallucinations and undertranslations better than existing evaluation methods.

Links:

- Code [https://go.umd.edu/simqa]
- Research Talk [https://youtu.be/rorRvvECvL4]

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

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

HyoJung Han
Computer Science
University of Maryland
hjhan@cs.umd.edu

Marine Carpuat
Computer Science
University of Maryland
marine@cs.umd.edu

Jordan Boyd-Graber
CS, UMIACS, iSchool, LCS
University of Maryland
jbg@umiacs.umd.edu

Abstract

Detractors of neural machine translation admit that while its translations are fluent, it sometimes gets key facts wrong. This is particularly important in simultaneous interpretation where translations have to be provided as fast as possible: before a sentence is complete. Yet, evaluations of simultaneous machine translation (SIMULMT) fail to capture if systems correctly translate the most salient elements of a question: people, places, and dates. To address this problem, we introduce a downstream word-by-word question answering evaluation task (SIMQA): given a source language question, translate the question word by word into the target language, and answer as soon as possible. SIMQA jointly measures whether the SIMULMT models translate the question quickly and accurately, and can reveal shortcomings in existing neural systems—hallucinating or omitting facts.

1 Introduction

Recent advances in simultaneous machine translation (SIMULMT) hold the promise of breaking language barriers by democratizing simultaneous interpretation, a demanding form of real-time translation which currently requires trained human experts (Grissom II et al., 2014; Cho and Esipova, 2016). However, progress in this field is hampered by evaluation challenges. Since SIMULMT systems must output translations in real time before the input sentence is complete, evaluation methods should account for the timeliness of outputs, in addition to capturing the dimensions of translation quality that also matter in traditional MT settings (Ma et al., 2019; Arivazhagan et al., 2019). Adequate evaluation on SIMULMT is critical to build models that help a user decide when to take the right action quickly and correctly in a multilingual situation. Since human evaluation is too costly (and slow) to guide system development, evaluation is currently limited to quantifying the trade-off between translation latency and quality, as measured by standard translation quality metrics (Ma et al., 2020a; Stewart et al., 2018) or further addition of time penalty (Grissom II et al., 2014). However, it remains unclear how to interpret such scores in practice. Metrics designed to assess the quality of full-input machine translation (MT) are not well-suited to evaluating partial translations, and can fail to capture salient errors. Measuring timeliness by the average rate by which the SIMULMT system lags behind an ideal synchronous translator does not tell us to what degree SIMULMT translations are useful for practical purposes.

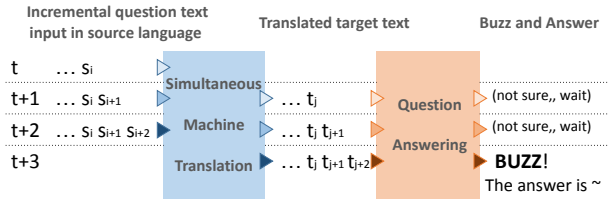


Figure 1: Overview of the Question Answering Evaluation for SIMULMT (SIMQA). Given a word-by-word QA task, the SIMULMT system produces translations of questions word by word. A good SIMULMT system will allow the downstream QA system to answer correctly as quickly as possible.

To address these issues, we propose to evaluate SIMULMT by measuring how well it helps execute a cross-lingual word-by-word question answering task (SIMQA). This task measures the quality of timely adequacy, or immediacy of adequacy, of SIMULMT more directly than existing evaluations. Given a source language question, the task consists of translating the question word-by-word into the target language, and answering it as soon as possible. This evaluation is inspired by the use of Question Answering (QA) to evaluate MT systems in full-input settings, where a correct answer is considered an indication that the salient source content was adequately translated (Tomita et al.,

1993; Sugiyama et al., 2015; Scarton and Specia, 2016; Krubiński et al., 2021). Unlike prior work, we evaluate translations on a word-by-word basis rather than waiting for a complete translation of the source, thereby evaluating whether source content is immediately conveyed adequately enough for a QA system to answer correctly.

We introduce a novel Cross-lingual Quizbowl Dataset XQB,[1] to systematically evaluate Polish-English and Spanish-English SIMULMT systems of varying latency, and compare them against full-input MT that translates complete inputs. By jointly accounting for timeliness and translation quality, the SIMQA evaluation reveals different trends that complement full-input MT and QA evaluations based on complete inputs, and also it can reveal critical SIMULMT errors by monitoring QA quality over time.

2 Motivation

2.1 SIMULMT Evaluation Challenges

Defining the quality of simultaneous interpretation is challenging. An early version of a practical guide to interpreters by the International Association of Conference Interpreters describes quality as "that elusive something which everyone recognizes but no one can successfully define" (AIIC, 1982). Grbić (2008) shows that there is no consensus in the definition of quality in the field of interpreting studies, where perspectives such as "quality as perfection" co-exist with "quality as fitness for purpose", where quality is viewed as satisfying user needs. Zwischenberger (2010) identifies more specific quality criteria that span three different dimensions: content (e.g., is the translation faithful, logical, complete?), form (e.g., is the translation grammatical? does it use the appropriate style and terminology?), and delivery (e.g., is the delivery synchronous, fluent, lively?). However, she finds that while interpreters all value faithfulness overall, the relative importance of these criteria varies depending on the type of interpretation assignment (e.g., technical congress vs. press conference) and the interpreter community surveyed.

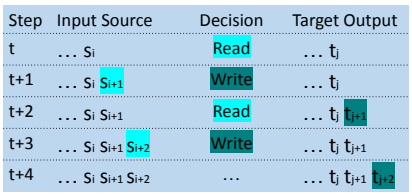
Additional issues arise when evaluating the quality of SIMULMT rather than human interpreting, since systems might err differently and cannot repair as humans do. The simultaneous nature of SIMULMT makes its evaluation more difficult than full-input MT, which is already recognized to be one of the hardest evaluations in natural language processing (Stewart et al., 2018; Cherry and Foster, 2019; Iranzo-Sánchez et al., 2021; Zhao et al., 2021). For instance, human experts agree that in simultaneous interpreting dropping extraneous information is sometimes acceptable, and that paraphrases that decrease the time between an utterance and its translation are not just acceptable but desired (He et al., 2016a). Given all of these difficulties, automatic evaluation for SIMULMT has either asked experts to evaluate translations along these dimensions (a high quality but slow and non-scalable solution) or adapted standard MT evaluations to minimize delay between utterance and translation. Current assessment of SIMULMT is thus simply based on two criteria: latency compared to an ideal synchronous interpreter and translation quality based on standard reference-based metrics designed for full-input MT. Also, the prior work of Grissom II et al. (2014) suggests Latency-BLEU (LBLEU) to jointly account for quality and 'expeditiousness' by adding a time penalty to BLEU. However, these metrics do not capture all the errors that neural systems are particularly prone to make, and do not tell us how useful the translations are.

2.2 Task-driven SIMULMT Evaluation

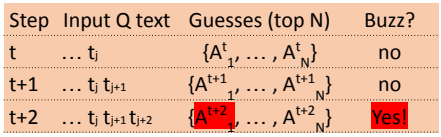
We take a different approach on evaluating SIMULMT and focus on task-driven evaluation. Imagine that you are transported to Warsaw, Poland and do not speak Polish. You are a contestant on a game show where you need to answer trivia questions. The questions give clues that get easier over time, and you can interrupt at any time to give your answer, with the goal to answer correctly before your opponent. Even though you know nothing about Polish, you get help from a SIMULMT system. Our hypothesis is that the higher the monolingual contestant's score on this game—where the latency of every word matters—is a reasonable proxy for how good the SIMULMT system is. As we discuss in Section 6, the quality of the underlying contestant (here a QA system) affects the score as well, but keeping it *constant* nevertheless makes it possible to compare SIMULMT systems.

While the specific scenario above might seem unlikely, this approach evaluates essential aspects of SIMULMT for a broad range of use cases. It can be seen as directly measuring the "fitness for purpose" of SIMULMT on a task where adequacy is particularly important, which aligns with impor-

[1] https://go.umd.edu/simqa



(a) Simultaneous Translation (SIMULMT) system in step-wise view



(b) Quizbowl Question Answering system in step-wise view

Figure 2: Overview of the element system of SIMQA, which is based on a pipeline of SIMULMT and QB.

tant dimensions of quality from interpreting studies discussed above. This setup targets error types that are likely to be particularly problematic with state-of-the-art neural MT systems.

3 Simultaneous Question Answering Evaluation Framework

In this section, we describe the two components of our framework, simultaneous translation (Section 3.1) and the word-by-word QA task (Section 3.2), and how they are integrated (Section 3.3).

3.1 Simultaneous Machine Translation

Simultaneous Machine Translation (SIMULMT) starts translating the prefix of an input sequence before the complete sequence is available, unlike traditional full-input MT which translates given a complete source sequence as an input. We use the term "full-input" MT to refer to translating entire sequences one at a time, which is sometimes also called "offline" MT in the literature.

The key design consideration for SIMULMT systems is therefore to choose the policy that governs whether to wait to receive more source tokens, or to write the output given the current prefix. Figure 2a illustrates this process. Among the many strategies that have been proposed in the literature (Cho and Esipova, 2016; Gu et al., 2017; Ma et al., 2019; Zheng et al., 2020b; Arivazhagan et al., 2019; Ma et al., 2020b), we choose WAIT-k (Ma et al., 2019) as a simultaneous translation model, because it is conceptually simple, relatively easy to implement, and achieves competitive results in existing evaluations. WAIT-k models simply wait until k source tokens have been produced, and then start to alternate writing target tokens and reading source tokens. When the model exhausts the source tokens, it continues to write until end of sequence.

For measuring latency, we use differentiable average lagging (Arivazhagan et al., 2019, DAL), which represents the average rate by which the SIMULMT system lags behind an ideal synchronous translator. Usually, DAL is k for WAIT-k SIMULMT models.

3.2 Word-by-Word Question Answering

We choose Quizbowl (Boyd-Graber et al., 2012, QB) as a proxy task for an evaluation of the simultaneous translation model since the task also deals with incremental inputs and with sequential decision making. QB is a trivia game where questions come in word-by-word: the questions get easier over time. Players of QB must interrupt the question and answer as soon as they know the answer (i.e., before their opponent). This problem has been investigated in monolingual settings in prior work. Here we consider a new *cross-lingual* version to evaluate a SIMULMT model.

As in simultaneous translation, the key trade-off in QB is accuracy vs. speed. After every input word, the system produces its top N guesses given the current input—which consists of the words used in the question in the monolingual case, and of the outputs of the SIMULMT system in the cross-lingual case. The QA system must decide whether to trust that partial translation, as is explicitly modeled in some simultaneous interpretation systems (Grissom II et al., 2014; Stewart et al., 2018).

Where the value of QA comes through is in the ease of evaluation: it is much easier to decide whether an answer is correct than to decide if a translation is "good". When humans compete on QB, it is typically head-to-head: on how many questions did Player A answer before Player B. While this metric has been used (He et al., 2016b), it is cumbersome. At the same time, while answering earlier is better than answering later, the number of characters that the system needs to answer is not an informative metric, as some parts of the question are more important than others.

As a result, we follow Rodriguez et al. (2019), who propose the expected wins (EW) metric to evaluate a system's ability to win at QB: EW represent the probability that a system would beat the average QB player on the question. Each question's score is the value of EW which is a function of rela-

tive position *only if* the prediction is correct, and 0 if the answer is wrong. Initially, the EW function is near 1.0—answering at the start of the question will always defeat an opponent—while at the end of the question it is near 0.0: most players will have answered the question correctly. In between, the function monotonically decreases, reflecting the average player's ability to answer a question. This function is trained on existing English questions over a variety of topics.[2]

3.3 SIMQA—QA Evaluation for SIMULMT

Task Overview. We define the task of Question Answering with SIMULMT (SIMQA) as follows: an input query text in the source language is fed into a SIMULMT model and the translated target query text is then processed with a QA model to produce an answer in the target language. Both systems process their respective inputs on a word-by-word basis, which results in the tightly integrated pipeline of Figure 1.

Concretely, this task is made possible by a cross-lingual Quizbowl dataset (XQB, described in Section 4), which provides parallel question-answer pairs in Polish-English and Spanish-English. The SIMULMT system translates the Polish question into English and an English QA model provides an answer. The QB buzzer model decides whether to buzz and return the answer, or to wait. The system answer is then compared to the English reference.

Integration Details. Since QB questions are long sequences that can include multiple sentences, we first split a paragraph of an input query (P_S) into unit sentences ($\mathbf{S} = \{S_1, ..., S_l, ...\}$), and then feed each source sentence token by token to the SIMULMT model ($S_l = \{s_1, ..., s_i, ...\}$). For each output time step ($T_l = \{t_1, ..., t_j, ...\}$), a generated token is fed into the word-by-word QA system. It decides whether to buzz and give its current answer, or whether to wait for more target tokens.

In the monolingual QB setting, the buzzing position is simply defined based on the current position in the input query. In the cross-lingual setting, we need to define the buzzing position τ_S on the source side in terms of target buzzing position τ_T as we now translate corresponding source text into target text.[3] The buzzing position in source text τ_S is:

$$\tau_S = \sum_{k=1}^{l-1} |S_k| + g_l(\tau_T - \sum_{k=1}^{l-1} |T_k|) \quad (1)$$

where $g_l(j)$ is delay at target position j: the number of source tokens read by the agent before writing the jth target token with a certain sentence l. In the case of WAIT-k model, $g_l(j)$ is approximately $min(k + j - 1, |S_l|)$.

This framework can also compare full-input MT on the QA task while ignoring incremental input. In the case of sentence translation for QB, we generalize the calculation of source buzz position by setting k to ∞.

# of Questions	# of Qs	# of Sentence
Source Questions ($\mathcal{S}$)	1,132	3456
English Ans matched ($\mathcal{M}$)	965	3095
Human Translated ($\mathcal{H}$)	512	1661
Guessable ($\mathcal{G}$)	341	1133
Answerable ($\mathcal{A}$)	314	987

Table 1: Statistics of XQB-*pl*. Usually one question consists of three or four sentences. The set $\mathcal{G}$ are questions that at least one model is able to guess correct answer in its top N guesses at any step. $\mathcal{A}$ are questions that at least one model guess correctly (top guess matched with the answer) at any step. In this paper, we refer to $\mathcal{H}$ as default set of XQB unless specified. ($\mathcal{A} \subseteq \mathcal{G} \subseteq \mathcal{H}$)

4 Experiment Settings

4.1 XQB : the Cross-lingual Quizbowl Dataset

Our SIMQA evaluation relies on QB questions written in languages other than English. We collect multilingual QB data in collaboration with International Academic Competitions (IAC).[4] These questions are used in competitions with grade school students in their local language. The domain questions include history, science and geology. The questions have a fixed distribution over subjects and difficulty levels to make it fun for many levels of contestants.

We construct the Cross-lingual Quizbowl test set (XQB) by translating Polish answers into English using language links across Wikipedia pages, augmented with the Wiki-Titles dataset from WMT.

[2] http://datasets.qanta.org

[3] To simulate a time-sensitive cross-lingual question answering environment, the buzzing position should be calculated on source sentence since the position is supposed to indicate where the whole QA system—including SIMULMT—outputs the prediction by consuming and processing the source question up to current point.

[4] https://www.iacompetitions.com

This yields a set of 965 Polish questions paired with English answers. In addition, we obtain reference translations for the Polish questions into English by asking human translators to post-edit full-input MT, providing randomly ordered Google Translate (GT), SMT and Transformer-big outputs (Section 4.2), to avoid biasing references toward a single MT system. This results into a parallel corpus of 512 questions and answers in Polish and English. An automatic sentence splitter splits questions into a single sentence for MT. The statistics of constructed XQB is detailed in Table 1.[5]

Following the same process, we collect 148 Spanish questions with English answers, which also come with human reference translations, as described in Appendix E.

4.2 Model Settings

Dataset. We use the Polish-English WMT2020 dataset (Barrault et al., 2020) to train all the MT models. Our MT models are trained only on the general domain corpus of WMT. They are *not* fine-tuned nor adapted to the XQB domain, which we treat exclusively as held-out evaluation data. Fast-align (Dyer et al., 2013) filters the training set, resulting in 8M sentences. We use the standard newsdev2020 and newstest2020 as development and test set. The Quizbowl system is the standard English training set from Rodriguez et al. (2019).

MT. The default setting for neural machine translation (NMT) and SIMULMT models follow the big configuration of transformer (Vaswani et al., 2017). The BPE shared vocabulary size is 40k (Sennrich et al., 2016). SIMULMT models are trained using WAIT-k (Ma et al., 2019), where $k \in \{3, 6, 9, 12, 15\}$, with a uni-directional encoder. The emission rate is set to 0.97 based on the tokenized training set. As a full-input baseline, we train a Transformer-big model ("Transf"). For statistical machine translation (SMT), we use the Moses toolkit (Koehn et al., 2007). The base NMT and SIMULMT models are trained up to 300k train steps on single A6000 GPU—each step is a batch of approximately 16384 tokens.[6] All models except for Google Translate, baseline, and SMT are trained three times and values in the results section includes means and standard deviations.

QA. QA is a GRU Guesser and a LSTM Buzzer.

[5]Example of questions can be found in Appendix A.

[6]An evaluation of all our MT systems on the publicly available WMT test set is available in Appendix B.

4.3 Evaluation Metrics

Extrinsic QB metrics. We compute the Expected Win (EW) scores, following Rodriguez et al. (2019). EW is the expected probability of winning against an average player as a function of buzzing position and empirically estimated based on the human gameplay. The EW probablity is only added when the buzzer buzzes for the first time and the answer is correct (as in an actual game). In this paper, however, there is only one player, so the score calculation under the competitive setting may be strict. Therefore, we also calculate EW with an oracle buzzer (EWO) where the buzzer would buzz as soon as the guesser gets the correct answer. As a result, EWO is a more lenient version of EW.

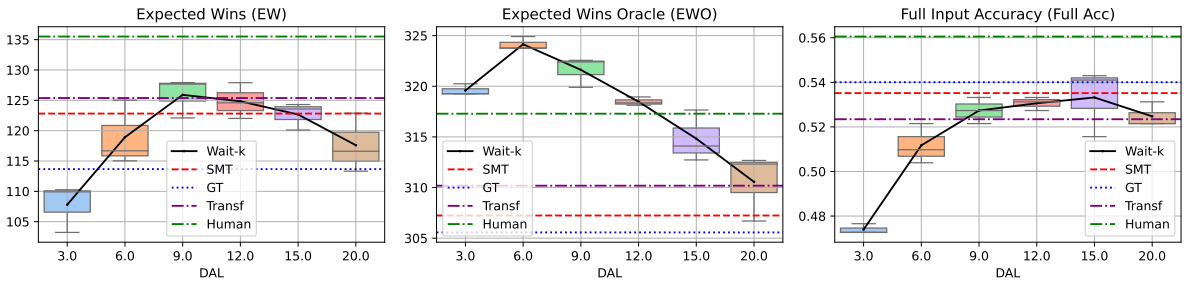
Intrinsic MT metrics. We compute BLEU (Papineni et al., 2002; Post, 2018), COMET (Rei et al., 2020), and BertScore (Zhang* et al., 2020).We use Differentiable Average Lagging (Arivazhagan et al., 2019, DAL) as a latency metric for SIMULMT.

5 Main Results

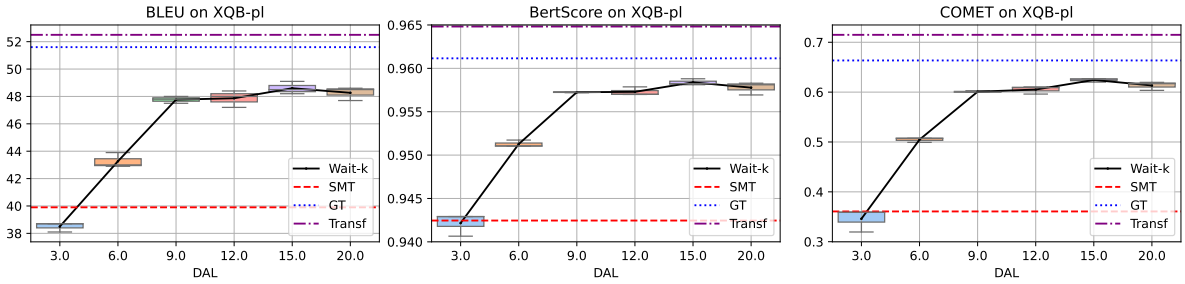
Figure 3 shows MT and QA metrics across MT systems on the Polish-English task.

MT Metrics. As expected, traditional metrics of MT quality all increase monotonically as the SIMULMT system waits longer before translating in Figure 3b. Translation quality increases steeply by 10 BLEU points from WAIT-3 to WAIT-9 and plateaus around 48 BLEU with longer wait times, 4 BLEU points below the translation quality of the full-input system that uses the same architecture and training data as SIMULMT ("Transf"). BertScore and COMET show consistent trends. The quality of SIMULMT models falls roughly between that of full-input SMT, which is close to or slightly above WAIT-3 depending on the metric, and that of full-input NMT systems, with our Transformer followed by Google Translate.

SIMQA Results. By jointly capturing how MT quality and latency impact the probability of winning, the QA evaluation in Figure 3a paints a different picture. While EW and EWO initially improve with longer latency, the translation quality gains have diminishing returns on QA as latency increases. The EW scores shows that the WAIT-9 models have better timely adequacy than WAIT-12, even though WAIT-12 models have access to three additional source tokens, and EW scores decrease



(a) Plots of {EW, EWO, Full-input Accuracy} vs DAL from QA metric.



(b) Plots of {BLEU, BertScore, COMET} vs DAL from MT metric.

Figure 3: Plots of QA and MT metric on XQB-*pl*. Different results on both metrics indicate that SIMQA metrics show the joint impact of adequacy and timeliness, and traditional SIMULMT does not fully assess the ability to convey core information in a timely manner.

further with higher latency. The EWO scores follow the same trend but peak with WAIT-6, indicating that timely adequacy could be obtained with a lower latency given an oracle buzzer.

Full-input MT. We compare SIMULMT systems with full-input MT. The scores of full-input MT systems are represented by horizontal lines in Figure 3. First, we compare the QA accuracy of these models based on the complete translation of the input question, thus ignoring the consideration of timeliness (Figure 3a). In this setting, SIMULMT systems are less accurate than full-input MT systems by at most 6 points, and the WAIT-15 setting is on par with full-input MT systems. While NMT outperforms SMT by a wide margin according to MT evaluation metrics, their accuracy is close on the QA task, falling roughly 2 points below the upper-bound achieved with human translation.

Next, we return to the SIMQA evaluation of timely adequacy via EW and EWO scores. We compute these scores for full-input MT models by assuming that they are WAIT-k models where k is ∞, and calculate the source buzzing position and corresponding EW score accordingly. In this setting, the best SIMULMT system (WAIT-9) outperforms full-input MT systems according to EW. Given an oracle buzzer (EWO), all WAIT-k models improve over full-input MT, and four of them even improve over human translation. While the accuracy of Transformer and SMT lag slightly behind that of Google Translate on full input accuracy, the Transformer and SMT systems are the best and second best of full-input MT according to EWO and EW scores, indicating that our local models' timely adequacy is higher than that of a commercial system.

Taken together, these results confirm the potential of SIMQA to evaluate the joint impact of adequacy and timeliness. This is further augmented by the results of the experiment with XQB-*es* which shows consistent trends in Appendix E. Design choices for SIMULMT systems, namely the WAIT-k policy, have a clear impact on SIMQA metrics. The discrepancy between the SIMQA and MT results confirms that traditional SIMULMT evaluation does not suffice to assess systems' ability to convey core source information correctly and quickly.

6 Consistency on multiple QA models

Section 5 only uses the RNN QA model. This consistency evaluates MT's effect on QA metrics. To see how the choice of QA system affects the results, we compare the RNN guesser with BERT (Devlin et al., 2018) and ElasticSearch (Gormley and Tong, 2015) guesser. This experiment uses the English answer matched set $\mathcal{M}$ in Table 1 to use all available

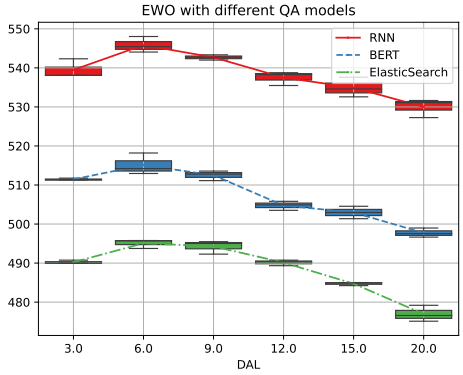


Figure 4: Plots of EWO with different QA models. The QA metric shows consistency across the various QA models, which confirms the soundness of our approach.

questions and report EWO for the WAIT-k models. While the RNN guesser outperforms BERT and ElasticSearch, EWO follows consistent trends with all guessers as DAL increases, with peak timely adequacy for WAIT-6. This confirms the soundness of our evaluation approach. The details of the experiment and full results of each QA metric are in Appendix C and Figure 7.

7 Step-wise Visualization of SIMULMT Errors

This section moves away from aggregate evaluations and analyzes the behavior of the SIMQA system step-by-step on a single question at a time. We show that monitoring the quality of question answering over time can reveal translation errors.

Method. Figures 5a and 5b show how the answers of the QA system and their quality change as the SIMQA system consumes the input question. The x-axis represents the relative position of a guess based on the question on the source side. The target word where the QB system generates a guess can be mapped to the location of the part of the source question that has been consumed by SIMQA. We normalize by the length of the source question to obtain the relative character position used on the x-axis of our plots. The y-axis uses mean reciprocal rank (MRR) to represent the goodness of the current answer. Reciprocal rank is the inverse of the position of the highest ranked answer, and if there are no correct answers within the top N then the reciprocal rank is 0. (higher is better) We use log MRR for more even plots. Here N is set to 50, and the lowest value of log MRR is -4. MRR can also be used for overall system evaluation, which results in similar trends as those described above with the EW and EWO metrics (refer to the Appendix D for details).

Hallucination. In the first example (Figure 5a and Table 2a), the WAIT-9 model is confused during the first half of the question and mostly guesses the correct answer ("Longitude") in the next half. The red curve, which plots a binary indicator for SIMULMT output words that are unaligned to the source,[7] shows that unaligned words often are critical MT errors and correlate with steep drops of log MRR.

At position (1), marked in Figure 5a and Table 2a, the Polish word "dwuścienny" is translated into a close but incorrect translation "double-wall", and this causes log MRR to drop slightly. At (2) and (3), the word "southern" is hallucinated by the SIMULMT model at each point, and there is no similar corresponding word in the given source. This hallucination causes log MRR to drop to zero. By contrast, at position (4), the QB system is robust enough to ignore the unaligned stopword "a".

BLEU does reveal translation errors, but it treats every word the same and is calculated after the full output, while SIMQA focuses on essential words that change the answer prediction in real-time. For example, when "półpłaszczyzna" is erroneously translated as "southern" instead of "half-plane" in Figure 5a, QA is misled and guesses wrong, but ignores minor errors such as excluding "the" or "a", while both cases have same impact on BLEU.

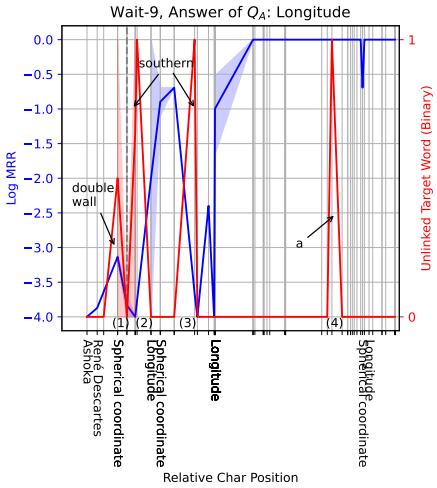
Under Translation. In Figure 5b, there is a flat span where log MRR dosen not change for more than half the question. This is because even though the WAIT-3 models consume source words continuously, it does not output any new words due to an under-translation error. As can be seen in Table 2b, the translation output of WAIT-3 is too short, and as a result the QA system generates a wrong guess.

These two examples illustrate that QA provides useful signals to pinpoint critical translation errors in SIMULMT outputs, and suggest that SIMQA might provide a strategy for more systematic error detection on the fly in future work.

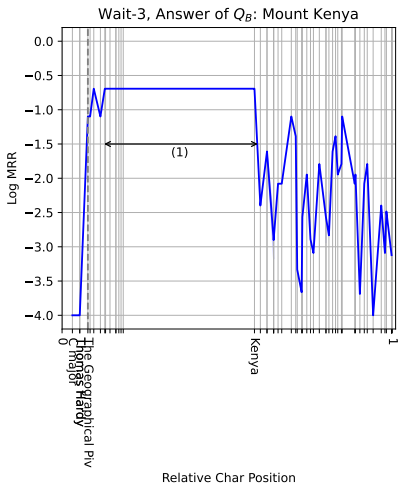
8 Related Work

QA as Evaluation. Answering questions provides a natural way to evaluate whether humans can comprehend MT output. For instance, Tomita

[7] With the fast-align model used to filter the training set.



(a) Example of translation error including hallucinations on the output of WAIT-9 model with log MRR changes. Unlinked Target word (red) is one if target word generated is unaligned with given source text at the position.



(b) Example of under-translation of WAIT-3 model with steady log MRR.

Figure 5: Step-wise analysis with relative character position vs log MRR plots with examples of translation errors. The x-axis indicates relative position of the question on the source side, and the changes of top guesses are presented on the relative position. The ticks in x-axis represents the point when the model generates guesses.

From	Time	Q_A Text	Guess	
Source	(1)-(3)	Tę współrzędną wyznacza kąt (1)dwuścienny między (2)półpłaszczyzną południka zerowego a (3)półpłaszczyzną południka ...		
Human	(1)-(3)	This coordinate is determined by the (1)dihedral angle found between the (2)half-plane prime meridian and the (3)half-plane meridian ...	Longitude	✓
WAIT-9	(1)	This coordinate determines the **double-wall**	Spherical coordinate	✗
WAIT-9	(2)	This coordinate determines the double-wall angle between the **southern**	Spherical coordinate	✗
WAIT-9	(3)	This coordinate determines the double-wall angle between the southern half of the meridian plane and the **southern**	Spherical coordinate	✗
Source	(4)	Tę miarę liczy się od południka zerowego (Greenwich) aż do południka 180 °. Aby otrzymać punkt, ...		
Human	(4)	This measure counts from the zero meridian (Greenwich) to the meridian 180 °. To get *a* point, ...	Longitude	✓
WAIT-9	(4)	... This measure counts from the south of the zero (Greenwich) to the south 180 °. To get *a* point	Longitude	✓

(a) Translation of WAIT-9 from source Q_A. The underlined text is incorrectly translated or hallucinated to boldface which leads to wrong prediction to the question.

From	Time	Q_B Text	Guess	
Source	(1)	W 1899 Halford Mackinder twierdził, że jest pierwszą osobą, która wspięła się na tę górę, a ludzie Kikuju nazywają tę górę Kirinyaga, co oznacza „tę ze strusiem ”, podczas gdy Masajowie wierzą, że ich przodkowie zeszli z tej góry na początku czasu.		
Human	(1)	In 1899, Halford Mackinder claimed that he was the first person to climb this mountain, and the people of Kikuyu named this mountain “Mount Kirinyaga”, meaning „the one with the ostrich”, while the people of Maasai believe that their ancestors descended from this mountain at the beginning of time.	Mount Kenya	✓
WAIT-3	(1)	In 1899, Halford Mackinder claimed that he was the first person to climb the mountain	Geographical Pivot of History	✗

(b) Translation of WAIT-3 from source Q_B. The underlined text is not translated by WAIT-3 which leads to a wrong guess.

Table 2: Translation Error examples that influence QA performance

et al. (1993) use a reading comprehension task from the TOEFL test to evaluate the quality of English to Japanese MT. Jones et al. (2005) find that an Arabic-to-English MT system made it possible for English speakers to pass Arabic Level 2 on a standard defense language proficiency test. Scarton and Specia (2016) use reading comprehension questions to obtain human assessments of MT at the document level. Forcada et al. (2018) find the use of gap-filling can be a reasonable alternative to reading comprehension questionnaires for ranking MT systems.

Automatic QA can evaluate whether full-input MT adequately conveys core elements of the source text on a larger scale. Some approaches rely on manually curated QA test beds (Sugiyama et al., 2015; Sun et al., 2020), while Krubiński et al. (2021) use QA generation to generate questions from references and extract an answer. While these papers show the promise of QA for evaluating MT, they focus on the full-input setting.

QA has also been used as an extrinsic evaluation for text summarization systems (Eyal et al., 2019; Wang et al., 2020; Durmus et al., 2020; Volpi and Malagò, 2020) to complement existing evaluation metrics, which are notoriously insensitive to factual inaccuracies and inconsistencies.

SIMULMT Evaluation. Most SIMULMT metrics focus on quantifying the timeliness of SIMULMT, using Average Proportion (Cho and Esipova, 2016, AP), Continuous Wait (Gu et al., 2017, CW), Average Lagging (Ma et al., 2019, AL), or Differentiable Average Lagging (Arivazhagan et al., 2019; Cherry and Foster, 2019, DAL). Though widely used as official metrics of IWSLT shared tasks (Ma et al., 2020a), these metrics are used at the sentence-level, which departs from realistic usage. Recent work improves latency measurement by adapting them to streaming input (Iranzo-Sánchez et al., 2021) and by introducing dedicated interpretation test sets (Zhao et al., 2021). However, all of these are used in combination with traditional MT metrics on complete outputs, which does not directly measure the impact of timely adequacy.

Adding a time penalty to BLEU could also jointly account for quality and latency in one metric. Grissom II et al. (2014) define the Latency-BLEU (LBLEU) to measure 'expeditiousness' and quality by calculating a word-by-word discrete integral across the input. However, LBLEU and related metrics retain all the weaknesses of BLEU. In particular, BLEU treats every word equally, and does not explicitly penalize mistranslating the most important words for answering a question. As a result, it does not directly align with users' information needs,[8] while our method can directly measure the impact of keyword (mis)translations and of translation delays.

9 Conclusion

The main motivation of simultaneous translation (SIMULMT) is to convey the core meaning in a source sentence as quickly as possible. However, current SIMULMT research measures the quality of output on "finalized" translations and separately considers latency measures, which cannot fully reflect the quality of timely adequacy.

This paper introduces a cross-lingual word-by-word question answering task (SIMQA) to quantify the timely adequacy of SIMULMT more directly. Our experiments on a cross-lingual Quizbowl task show that some WAIT-k SIMULMT systems can win more often than full-input MT systems with higher BLEU score, and can outperform full-input human translations with an oracle buzzer. Overall our SIMQA results complement intrinsic QA and MT metrics by jointly accounting for timeliness and translation quality, and suggest that SIMQA can diagnose critical SIMULMT errors on the fly.

These results represent a first step toward broadening the scope of SIMULMT evaluation to more directly assess its usefulness in specific scenarios. In future work, we would like to evaluate on more languages and investigate the impact of varying each component of the SIMQA pipeline, including using alternate SIMULMT architectures, and providing SIMULMT outputs to human contestants in addition to automatic QA systems. Finally, we would like to expand the modality to speech for more realistic SIMULMT and QA.

Limitations

The experiments in this paper are limited to European languages, and to one direction, into English. Polish is a West Slavic language, which differs from English in several ways: it is a highly fusional language which has seven grammatical cases. It has relatively free word order, although it often follows an SVO structure. Therefore, we can expect some reordering to be needed but probably not as

[8] Furthermore, in our experiments, LBLEU increases monotonically with latency, adding no information to BLEU.

much as when translating from pairs with significantly different word order like Korean or Japanese into English. We also experiment with Spanish, which requires minimal word reordering into English. Given that monotonicity between source and target language has an impact on the performance of SIMULMT (Chen et al., 2021; Han et al., 2021), experimental results in languages with different word order could be different.

However, we expect SIMQA to remain an insightful evaluation tool for languages with largely different word orders, as it will capture delayed translations from SIMULMT. For example, in a question with "married" ending a long sentence, the QA system may fail to answer because it lacks the essential relationship between entities. On the other hand, conventional BLEU would consider this omission equivalent to any other word. Therefore, we expect SIMQA to be no worse on different word orders.

Some evaluation settings were directly borrowed from English Quizbowl for our cross-lingual version of the game. Specifically, the Expected Wins function (EW) is trained on English Quizbowl questions. However, since the Polish version of the game is directly modeled after the English game, we do not expect this to be an issue.

More importantly, our experiments are all simulations. It remains to be seen how SIMULMT systems would fare when used by human contestants rather than QA systems, or when pitted against bilingual or monolingual human contestants, possibly assisted by human simultaneous interpreters. This work relies on human reference translations (obtained under the full-input setting) as a first step, and did not compare against simultaneous interpretation by humans.

Finally, our work only considers the text modality due to limited resources. However, the most practical modality of SIMULMT and QB for that matter is speech.

Ethics Statement

As mentioned in 4.1, the Multilingual QB questions are established by an international academic organization, and we are using these questions with their explicit permission. The questions are in the public domain.

The workers who translated or post-edited Quizbowl questions were paid at a rate of USD 0.5 per question. The resulting estimated hourly wage is above the minimum wage per hour in the United States.

Acknowledgements

We thank International Academic Competitions for creating and providing multilingual Quizbowl questions. Specifically, we would like to thank David Madden (Executive Director), Lee Holden (Senior Director of Question Production), Natalia Stasik (Director, IAC Polska), Franek Alverado (Director of Question Production—IAC Polska), Gabriel Roldos (Director, IAC Ecuador), and Saad Bashir (Director of Question Production—IAC Ecuador) for their valuable support. We thank the anonymous reviewers, Pedro Rodriguez, Shi Feng, Weijia Xu, Elijah Rippeth, Aquia Richburg and the members of the CLIP lab at UMD for their insightful and constructive feedback. This work is supported by NSF Grants IIS-1822494 and Grant IIS-1750695 and by ODNI, IARPA, via the BETTER Program contract #2019-19051600005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the US Government. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

AIIC. 1982. Practical Guide for Professional Interpreters. Technical report, Geneva, Switzerland.

Ashkan Alinejad, Maryam Siahbani, and Anoop Sarkar. 2018. Prediction improves simultaneous neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3022–3027, Brussels, Belgium. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323, Florence, Italy. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, and George Foster. 2020. Re-translation versus streaming for simultaneous translation. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 220–227, Online. Association for Computational Linguistics.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual Open-Retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Jordan Boyd-Graber, Brianna Satinoff, He He, and Hal Daumé III. 2012. Besting the quiz master: Crowdsourcing incremental classification games. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1290–1301, Jeju Island, Korea. Association for Computational Linguistics.

Ozan Caglayan, Julia Ive, Veneta Haralampieva, Pranava Madhyastha, Loïc Barrault, and Lucia Specia. 2020. Simultaneous machine translation with visual context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2350–2361, Online. Association for Computational Linguistics.

Junkun Chen, Renjie Zheng, Atsuhito Kita, Mingbo Ma, and Liang Huang. 2021. Improving simultaneous translation by incorporating pseudo-references with fewer reorderings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5857–5864, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colin Cherry and George Foster. 2019. Thinking slow about latency evaluation for simultaneous machine translation. *arXiv preprint arXiv:1906.00048*.

Kyunghyun Cho and Masha Esipova. 2016. Can neural machine translation do simultaneous translation? *arXiv preprint arXiv:1606.02012*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Mikel L. Forcada, Carolina Scarton, Lucia Specia, Barry Haddow, and Alexandra Birch. 2018. Exploring gap filling as a cheaper alternative to reading comprehension questionnaires when evaluating machine translation for gisting. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 192–203, Brussels, Belgium. Association for Computational Linguistics.

Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The definitive guide: A distributed real-time search and analytics engine*. " O'Reilly Media, Inc.".

Nadja Grbić. 2008. Constructing interpreting quality. *Interpreting*, 10(2):232–257.

Alvin Grissom II, He He, Jordan Boyd-Graber, John Morgan, and Hal Daumé III. 2014. Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1342–1352, Doha, Qatar. Association for Computational Linguistics.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. 2017. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062, Valencia, Spain. Association for Computational Linguistics.

HyoJung Han, Seokchan Ahn, Yoonjung Choi, Insoo Chung, Sangha Kim, and Kyunghyun Cho. 2021. Monotonic simultaneous translation with chunk-wise reordering and refinement. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1110–1123, Online. Association for Computational Linguistics.

Hyojung Han, Sathish Indurthi, Mohd Abbas Zaidi, Nikhil Kumar Lakumarapu, Beomseok Lee, Sangha Kim, Chanwoo Kim, and Inchul Hwang. 2020. Faster re-translation using non-autoregressive model for simultaneous neural machine translation. *arXiv preprint arXiv:2012.14681*.

He He, Jordan Boyd-Graber, and Hal Daumé III. 2016a. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 971–976, San Diego, California. Association for Computational Linguistics.

He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé, III. 2016b. Opponent modeling in deep reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1804–1813, New York, New York, USA. PMLR.

Javier Iranzo-Sánchez, Jorge Civera Saiz, and Alfons Juan. 2021. Stream-level latency evaluation for simultaneous machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 664–670, Punta Cana, Dominican Republic. Association for Computational Linguistics.

D. Jones, E. Gibson, W. Shen, N. Granoien, M. Herzog, D. Reynolds, and C. Weinstein. 2005. Measuring human readability of machine generated text: Three case studies in speech recognition and machine translation. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 5, pages v/1009–v/1012 Vol. 5.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021. Just ask! evaluating machine translation by asking and answering questions. In *Proceedings of the Sixth Conference on Machine Translation*, pages 495–506, Online. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Xutai Ma, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. 2020a. SIMULEVAL: An evaluation toolkit for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 144–150, Online. Association for Computational Linguistics.

Xutai Ma, Juan Miguel Pino, James Cross, Liezl Puzon, and Jiatao Gu. 2020b. Monotonic multihead attention. In *International Conference on Learning Representations*.

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. 2018. Low-latency neural speech translation. In *Proc. Interspeech 2018*, pages 1293–1297.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan L. Boyd-Graber. 2019. Quizbowl: The case for incremental question answering. *CoRR*, abs/1904.04792.

Carolina Scarton and Lucia Specia. 2016. A reading comprehension corpus for machine translation evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3652–3658, Portorož, Slovenia. European Language Resources Association (ELRA).

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Craig Stewart, Nikolai Vogler, Junjie Hu, Jordan Boyd-Graber, and Graham Neubig. 2018. Automatic estimation of simultaneous interpreter performance. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 662–666, Melbourne, Australia. Association for Computational Linguistics.

Kyoshiro Sugiyama, Masahiro Mizukami, Graham Neubig, Koichiro Yoshino, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2015. An investigation of machine translation evaluation metrics in cross-lingual question answering. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 442–449, Lisbon, Portugal. Association for Computational Linguistics.

Shuo Sun, Suzanna Sia, and Kevin Duh. 2020. CLIReval: Evaluating machine translation as a cross-lingual information retrieval task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 134–141, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Masaru Tomita, Masako Shirai, Junya Tsutsumi, Miki Matsumura, and Yuki. 1993. Evaluation of MT systems by TOEFL. In *Proceedings of the Fifth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Kyoto, Japan.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Riccardo Volpi and Luigi Malagò. 2020. Evaluating natural alpha embeddings on intrinsic and extrinsic tasks. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 61–71, Online. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2020. Learning adaptive segmentation policy for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2280–2289, Online. Association for Computational Linguistics.

Shaolei Zhang and Yang Feng. 2021. Universal simultaneous machine translation with mixture-of-experts wait-k policy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7306–7317, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jinming Zhao, Philip Arthur, Gholamreza Haffari, Trevor Cohn, and Ehsan Shareghi. 2021. It is not as good as you think! evaluating simultaneous machine translation on interpretation data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6707–6715, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Baigong Zheng, Kaibo Liu, Renjie Zheng, Mingbo Ma, Hairong Liu, and Liang Huang. 2020a. Simultaneous translation policies: From fixed to adaptive. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2847–2853, Online. Association for Computational Linguistics.

Baigong Zheng, Renjie Zheng, Mingbo Ma, and Liang Huang. 2019. Simultaneous translation with flexible policy via restricted imitation learning. In *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, pages 5816–5822, Florence, Italy. Association for Computational Linguistics.

Renjie Zheng, Mingbo Ma, Baigong Zheng, Kaibo Liu, and Liang Huang. 2020b. Opportunistic decoding with timely correction for simultaneous translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 437–442, Online. Association for Computational Linguistics.

Cornelia Zwischenberger. 2010. Quality criteria in simultaneous interpreting: An international vs. a national view. *The Interpreters' Newsletter*, 15:127–142.

A Examples of XQB

Table 5 shows example questions of XQB-*pl* including full text of source question, human translation, and MT result of WAIT-3 & WAIT-6.

B Additional MT Metrics

Figure 9 show additional MT evaluation metrics of YiSi-1(Lo, 2019), Prism (Thompson and Post, 2020), and BLEURT (Sellam et al., 2020). We also provide the performance of all the models on standard WMT test set in Figure 8.

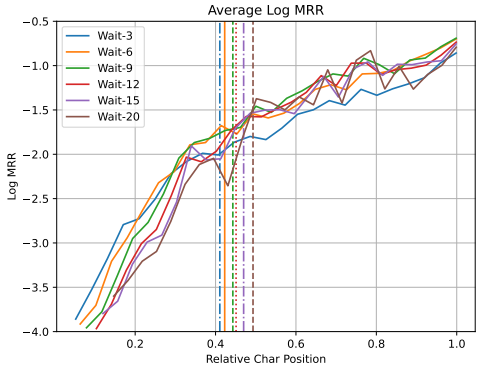


Figure 6: Average log MRR of all guessable set of questions on relative character position for each models. The vertical lines average position of buzzing point of each models. WAIT-9 shows generally good performance over other models.

C Results and Details of multiple QA models

In this section, we describe the details and show the full results of the experiment of different QA guesser models mentioned in Section 6. For additional QA models, we use BERT (Devlin et al., 2018) as Transformer model and ElasticSearch (Gormley and Tong, 2015) as Information Retrieval model. Both guessers are trained and indexed with same data used in the GRU guesser. The results on EW, EWO and full-input accuracy are shown in Figure 7. The trends for the EW metrics are not as consistent, which we attribute to the fact that EW is based on fewer data points than EWO.

D Average Mean Reciprocal Rank Evaluation

We show overall performance on a set of guessable questions ($\mathcal{G}$ in Table 1) for each model in Figure 6. The relative character position is binned into 25 section and y-axis is average of log MRRs of each bins for all guessable questions. As relative character position move toward one, or as the SIMQA consumes more information, log MRR increases to zero which means a model get closer to the answer. In Figure 6, WAIT-9 show generally top performance except for first one third part, which is consistent with our main result in Figure 3a. Since WAIT-k with higher k is closer to full-input MT, the MRR scores is usually positioned around the ending position of each sentence, and this causes the fluctuation of WAIT-12 $\sim$ model.

E Spanish XQB and other languages

E.1 XQB-es

In this section, we describe newly added Spanish set, XQB-es and its results on QA and MT metrics. We collect small set of Spanish questions from local competitions in Ecuador as presented in Table 3. We use Spanish-English WMT2013 (Bojar et al., 2013) dataset to train all the MT models. We use standard set of newstest2012, newstest2013 as dev and test set. We run experiments with same settings on XQB-pl except that we remove the ElasticSearch result due to severely low performance.

In Figure 10, EWO present similar trends across QA models as well as languages where the peak is mostly on $k = 6$ and the performance diminishes as latency increases. EW with RNN models shows similar trends with other languages while BERT

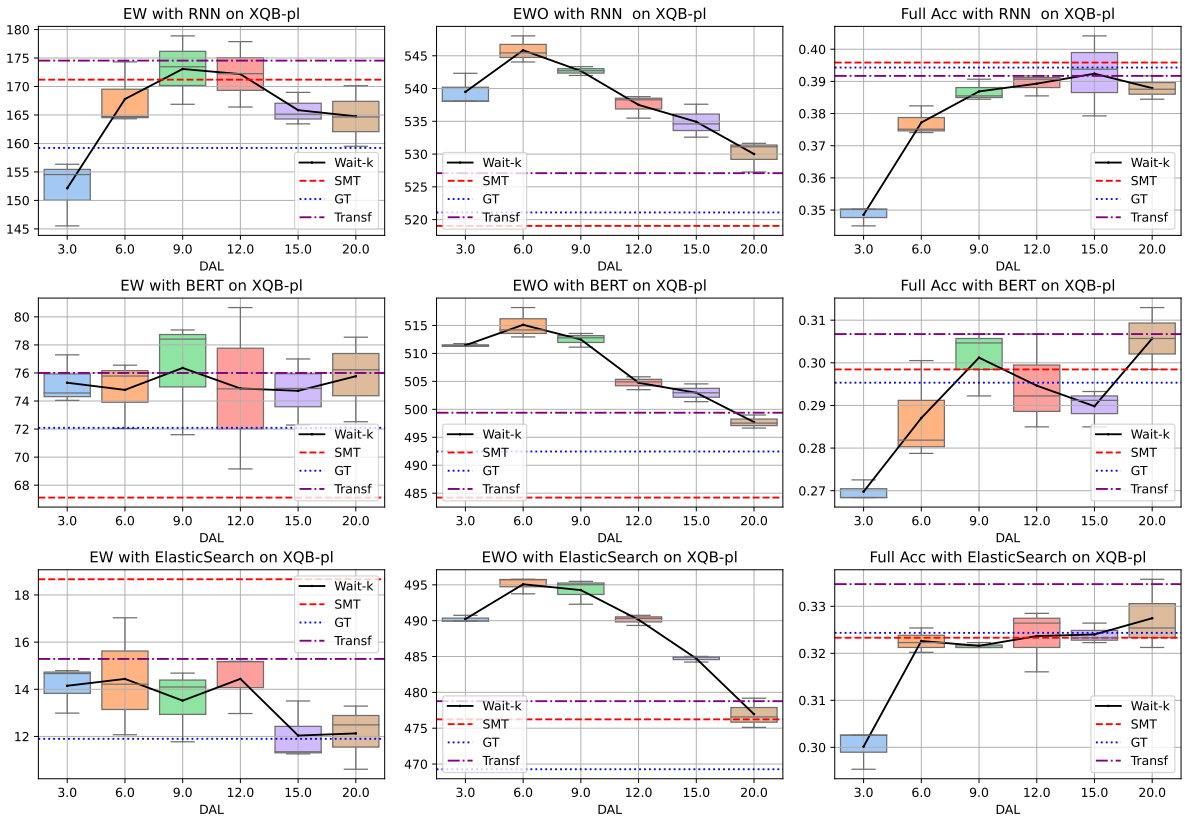


Figure 7: Plots of {EW, EWO, Full-input Accuracy} vs DAL from various QA models on English answer matched set $\mathcal{M}$ of XQB-*pl*.

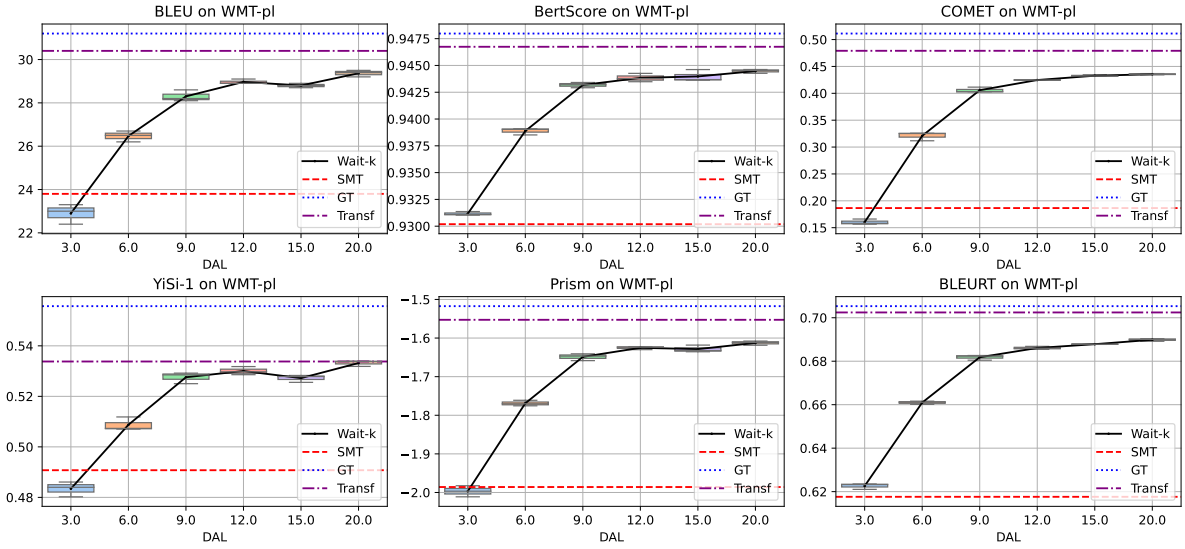


Figure 8: Plots of all evaluation metrics vs DAL from MT on WMT test set in Polish.

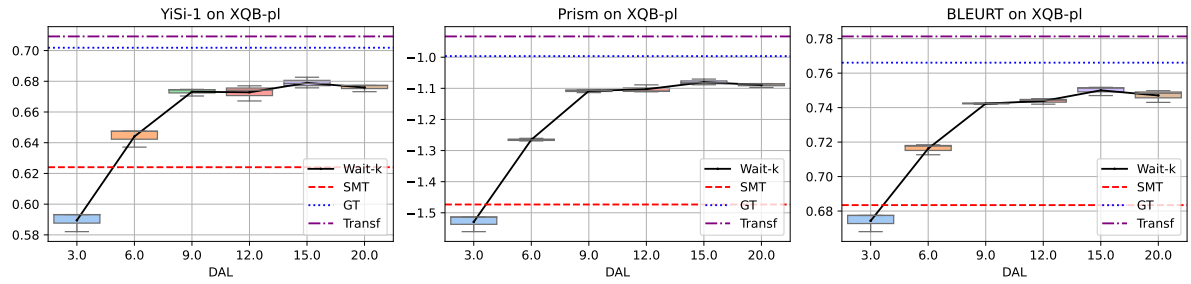


Figure 9: Plots of {YiSi-1, Prism, BLEURT} vs DAL from MT metric.

has different conclusion compare to Polish. We hypothesise it is due to high impact of noisy example among small set and lower full-input accuracy on SIMULMT models with high latency. Overall, we observe similar trends on SIMQA results with XQB-*es* compared to XQB-*pl* which indicates robustness of our SIMQA task.

E.2 Collection of XQB

We collect multilingual QA data in collaboration with an International Academic Competitions via local competitions in each country. Therefore, the creation of XQB is a gradual procedure as local competitions progress rather than a one-time mass production, and the ease of collection for new languages depends on competition popularity in a language. However, since QA competitions are active and the dataset is for evaluation (not for training), we can collect enough questions to SIMULMT models.

E.3 Overlaps across languages

For Polish, there are about 65% overlap between the Qanta English QB (Rodriguez et al., 2019) and XQB-*pl* answers. "France" and "sun" are the most frequent common answers, while "Białowieża Forest" or "Jan Brzechwa"(Polish poet) are in Polish only. For Spanish, the overlap is 64%, while the most common answer is "Spain", and "Julio Jaramillo" (Ecuadorian singer) as an example of Spanish only answer.

F Additional Related Works

Alongside the evaluation methods of latency, various SIMULMT models is also propose to optimally achieve faster translations with better accuracy. SIMULMT models can be categorized by the kind of policy. The SIMULMT model with fixed policy generates translation based on pre-defined policy without considering current status of the translation process. For example, the WAIT-k model proposed by Ma et al. (2019) waits for k tokens and alternates READ and WRITE. Even though this deterministic feature in the policy results in anticipation error, the quality is competitive with not-to-small ks, and faster speed on making decision of action is one of the strong point with the easy implementation as well. Many variations of WAIT-k have been proposed to address the shortcomings. For example, Caglayan et al. (2020) exploits additional visual context to complement missing source information. Finally, Zheng et al. (2020a) extended the WAIT-k to an adaptive policy by integrating a set of fixed policies models. Also, Zhang and Feng (2021) propose universal SIMULMT model with mixture-of-experts WAIT-k to achieve optimal performance under arbitrary latency.

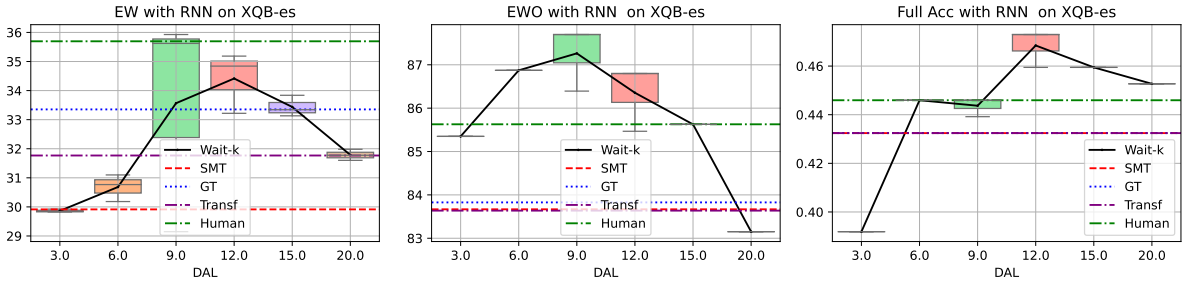
Many adaptive policies have also been suggested (Cho and Esipova, 2016; Gu et al., 2017; Zheng et al., 2019; Arivazhagan et al., 2019; Ma et al., 2020b). Cho and Esipova (2016) suggest to use greedy decoding to decide an action to write or read, while Gu et al. (2017) utilize a reinforcement learning to train agent with the objective of maximizing quality and minimizing latency. Advancing this work, Alinejad et al. (2018) adds a new action called PREDICT that anticipates upcoming source words.

Beside the kind of policy, availability of revision could be another characteristics of SIMULMT. While streaming approach only appends the generated tokens, re-translation allows limited revisions to the already presented partial translation (Niehues et al., 2018; Arivazhagan et al., 2020; Han et al., 2020). Re-translation can provide more accurate translation via correction of prior mistakes, however frequent revision can harm the user experience and may not suitable for speech applications.

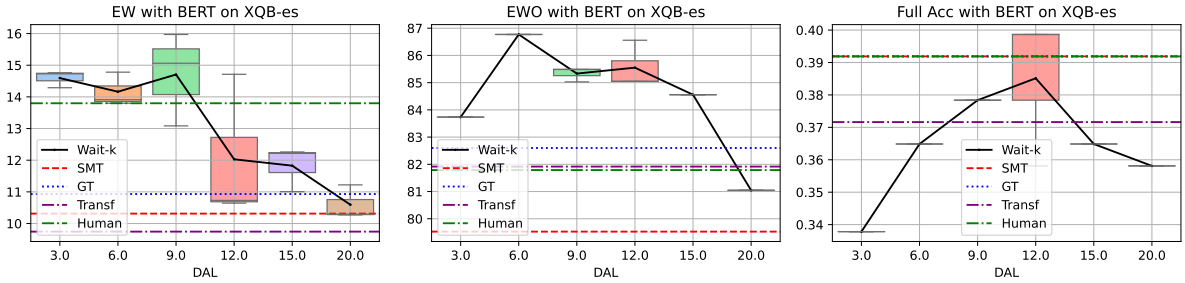
Recently, Arivazhagan et al. (2019) utilize hard attention for decision making and introduced DAL

# of Questions	# of Qs	# of Sentence
Source Questions ($\mathcal{S}$)	160	652
English Ans matched ($\mathcal{M}$)	148	603
Human Translated ($\mathcal{H}$)	148	603

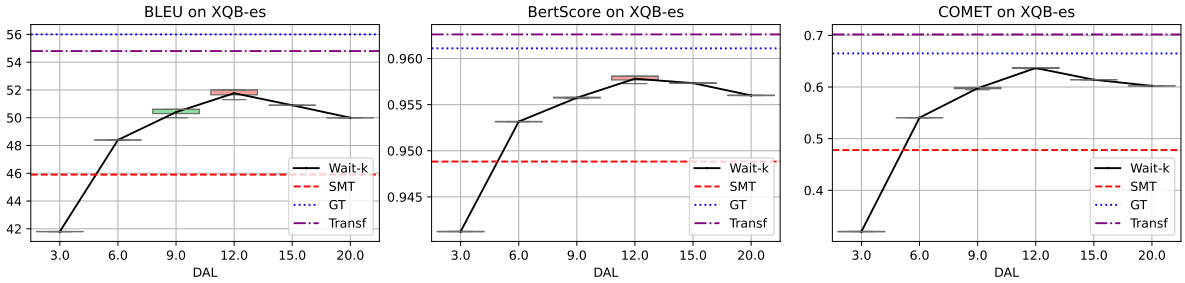
Table 3: Statistics of XQB-es.



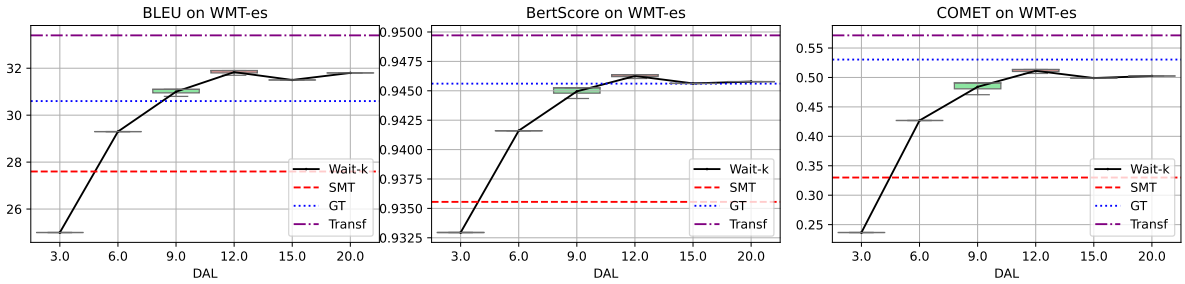
(a) Plots of {EW, EWO, Full-input Accuracy} vs DAL from QA metric with RNN on XQB-es.



(b) Plots of {EW, EWO, Full-input Accuracy} vs DAL from QA metric with BERT on XQB-es.



(c) Plots of {BLEU, BertScore, COMET} vs DAL from MT metric on XQB-es.



(d) Plots of {BLEU, BertScore, COMET} vs DAL from MT metric on Spanish WMT testset.

Figure 10: Plots of QA and MT metric on Spanish. Despite differences on language and size of set, SIMQA results on XQB-es has similar trends with XQB-pl, which indicates the robustness of the suggested evaluation task.

	QA Metrics		MT Metrics			
Russian	F1	EM	BLEU	COMET	BLEURT	Prism
Human	37	29.4	_	_	_	_
SMT	**24.43**	**18.82**	**25.7**	0.13	0.63	-1.90
m2m_418M	**24.26**	**17.65**	**36.1**	0.52	0.73	-1.64
m2m_1.2B	26.31	19.61	40.67	0.63	0.77	-1.33
m2m_12B	26.76	20.78	39.92	0.58	0.76	-1.44
Finnish	F1	EM	BLEU	COMET	BLEURT	Prism
Human	34.8	26.00	_	_	_	_
SMT	25.02	17.50	19.10	0.04	0.59	-2.30
m2m_418M	25.08	18.06	33.97	0.45	0.67	-1.27
m2m_1.2B	29.52	20.56	41.20	0.64	0.73	-1.10
m2m_12B	29.84	21.10	39.87	0.59	0.71	-1.18

Table 4: Comparison between MT models and several metrics in Russian and Finnish with XOR-TyDi QA dev set.

which is a differentiable version of average lagging (Ma et al., 2019, AL), in order to integrate latency measures into training losses. Ma et al. (2020b) incorporate this work into the multi-headed Transformer model. Furthermore, Zhang et al. (2020) proposes learning method to segment source input corresponds to possible target output.

G Full-input QA and MT Evaluation

Alternatively, we conduct experiments on non-dynamic QA system where complete and non-incremental static input generates one time output to verify our findings from QB in more general frameworks where the feature of dynamic inputs and decision making is removed. In this case, the query is a single interrogative question sentence. We choose to set up the experiment to translate only a question query for simplicity rather than translating both the query and the given passages.

For QA metrics, standard Exact Match (EM) and F1-Score are used. We use a multilingual development set of XOR-TyDi QA data (Asai et al., 2021). Among seven languages of XOR-TyDi, we choose Finnish (fi) and Russian (ru) for the experiments due to the availability of MT model pairs. For the QA model, we use a pre-trained DPR[9]model. For MT models, we use SMT, Google Translate (GT) and M2M (Fan et al., 2021) models in different parameter size—m2m100_418M (small), m2m100_1.2B (medium), m2m100-12B-avg-5-ckpt (large).[10] Since human translations are only given on training set of XOR, we randomly choose 2000 source-reference pairs from training set and use it as evaluating MT scores. The value of human in Table 4 is taken from Asai et al. (2021).

In Table 4, we evaluate the quality of full-input MT with XOR-TyDi dataset. Our experiment includes Russian and Finnish languages, and we use m2m model in different size and a SMT model. The findings in this setup is not exactly same as the QA performance of SMT translated questions is not the best or second best compare to other full-input MT models. However, we can observe that QA evaluation of SMT model still the better or similar compare to that of m2m_418M model while those two model shows large gap in MT evaluations. Also, m2m_1.2B and m2m_12B shows opposite measurement in QA and MT metrics, and it is interesting to see such disagreement is consistent across languages. While m2m_1.2B constantly outperforms m2m_12B in all MT metrics, m2m_12B is outperforming m2m_1.2B in extrinsic QA metrics. This indicates that QA and MT metrics show clear disagreement in the range of higher quality and in the range of lower quality as well.

[9] https://github.com/AkariAsai/XORQA/tree/main/baselines/DPR

[10] https://huggingface.co/facebook

Q_A	Text
Source	Tę współrzędną wyznacza kąt dwuścienny między półpłaszczyzną południka zerowego a półpłaszczyzną południka przechodzącego przez określony punkt na powierzchni Ziemi. Tę miarę liczy się od południka zerowego (Greenwich) aż do południka 180°. Aby otrzymać punkt, nazwij tę długość, która może przyjmować miary od 0° do 180° i może być wschodnia lub zachodnia.
Answer-src	Długość geograficzna
Human	This coordinate is determined by the dihedral angle found between the half-plane prime meridian and the half-plane meridian which passed through a specific point on the Earth's surface. This measure is counted from the prime meridian (Greenwich) to the 180° meridian. To gain a point, give the name of the length, which can measure between 0° to 180° and can be east or west.
Answer-tgt	Longitude
WAIT-9	This coordinate determines the double-wall angle between the southern half of the meridian plane and the southern half-plane passing through a certain point on the surface of the Earth. This measure counts from the south of the zero (Greenwich) to the south 180 °. To get a point, name the length that can take measures from 0 ° to 180 ° and can be eastern or western.
Q_B	Text
Source	W 1899 Halford Mackinder twierdził, że jest pierwszą osobą, która wspięła się na tę górę, a ludzie Kikuju nazywają tę górę Kirinyaga , co oznacza „tę ze strusiem", podczas gdy Masajowie wierzą, że ich przodkowie zeszli z tej góry na początku czasu. Góra ta jest głównym obszarem zlewiska rzeki Tana, największej rzeki w kraju, która może być również nazywana tak samo jak ta góra. Aby zdobyć punkt, podaj nazwę tej drugiej najwyższej góry w Afryce.
Answer-src	Mount Kenia
Human	In 1899, Halford Mackinder claimed that he was the first person to climb this mountain, and the people of Kikuyu named this mountain "Mount Kirinyaga", meaning „the one with the ostrich", while the people of Maasai believe that their ancestors descended from this mountain at the beginning of time. This mountain is the main catchment area of the Tana river, the largest river in the country, which is sometimes refered to by the same name as the mountain. To gain a point, enter the name of the second highest mountain in Africa.
Answer-tgt	Mount Kenya
WAIT-3	In 1899, Halford Mackinder claimed that he was the first person to climb the mountain. This mountain is the main area of the Tana River, the largest river in the country that can also be called the same as the mountain. To get a point, please specify the name of the other top in Africa.

Table 5: Example questions set of XQB-*pl*.