Automatic Estimation of Simultaneous Interpreter Performance

Craig Stewart,[1] **Nikolai Vogler,**[1] **Junjie Hu,**[1] **Jordan Boyd-Graber,**[2] **Graham Neubig**[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]CS, iSchool, UMIACS, LSC, University of Maryland

Abstract

Simultaneous interpretation, translation of the spoken word in real-time, is both highly challenging and physically demanding. Methods to predict interpreter confidence and the adequacy of the interpreted message have a number of potential applications, such as in computer-assisted interpretation interfaces or pedagogical tools. We propose the task of predicting simultaneous interpreter performance by building on existing methodology for quality estimation (QE) of machine translation output. In experiments over five settings in three language pairs, we extend a QE pipeline to estimate interpreter performance (as approximated by the METEOR evaluation metric) and propose novel features reflecting interpretation strategy and evaluation measures that further improve prediction accuracy.[1]

1 Introduction

Simultaneous Interpretation (SI) is an inherently difficult task that carries significant cognitive and attentional burdens. The role of the simultaneous interpreter is to accurately render the source speech in a given target language in a timely and precise manner. Interpreters employ a range of strategies, including generalization and summarization, to convey the source message as efficiently and reliably as possible (He et al., 2016). Unfortunately, the interpreter is pitched against the limits of human memory and stamina, and after only minutes of interpreting, the number of errors made by an interpreter begins to increase exponentially (Moser-Mercer et al., 1998).

---

[1]https://github.com/craigastewart/qe_sim_interp

(a) Standard simultaneous interpretation (SI)

(b) Computer assisted interpretation (CAI)
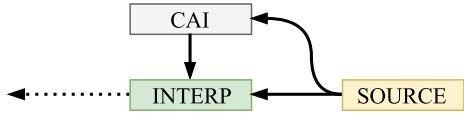
(c) Proposed CAI moderated by quality estimation
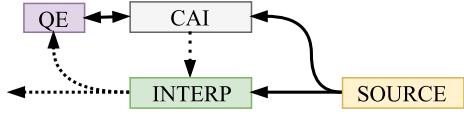
Figure 1: Simultaneous interpretation scenarios

We examine the task of *estimating simultaneous interpreter performance*: automatically predicting when interpreters are interpreting smoothly and when they are struggling. This has several immediate potential applications, one of which being in Computer-Assisted Interpretation (CAI). CAI is quickly gaining traction in the interpreting community, with software products such as Interpret-Bank (Fantinouli, 2016) deployed in interpreting booths to provide live and interactive terminology support. Figure 1(b) shows how this might work; both the interpreter and the CAI system receive the source message and the system displays assistive information in the form of terminology and informational support.

While this might improve the quality of interpreter output, there is a danger that these systems will provide *too much* information and increase the cognitive load imposed upon the interpreter (Fantinouli, 2018). Intuitively, the ideal level of support depends on current interpreter performance. The system can minimize distraction by providing assistance only when an interpreter is struggling. This level of support could be moderated appropriately if interpreter performance can be accurately predicted. Figure 1(c) demonstrates

how our proposed quality estimation (QE) system receives and evaluates interpreter output, allowing the CAI system to appropriately lower the amount of information passed to the interpreter, maximizing the quality of interpreter output.

As a concrete method for estimating interpreter performance, we turn to existing work on QE for machine translation (MT) systems (Specia et al., 2010, 2015), which takes in the source sentence and MT-generated outputs and estimates a measure of quality. In doing so, we arrive at two natural research questions:

1. Do existing methods for performing QE on MT output also allow for accurate estimation of interpreter performance, despite the inherent differences between MT and SI?
2. What unique aspects of the problem of interpreter performance estimation, such as the availability of prosody and other linguistic cues, can be exploited to further improve the accuracy of our predictions?

The remainder of the paper describes methods and experiments on English-Japanese (EN-JA), English-French (EN-FR), and English-Italian (EN-IT) interpretation data attempting to answer these questions.

2 Quality Estimation

Blatz et al. (2004) first proposed the problem of measuring the quality of MT output as a prediction task, given that existing metrics such as BLEU (Papineni et al., 2002) rely on the availability of reference translations to evaluate MT output quality, which aren't always available. As such, QE has since received widespread attention in the MT community and since 2012 has been included as a task in the Workshop on Statistical Machine Translation (Callison-Burch et al., 2012), using approaches ranging from linear classifiers (Ueffing and Ney, 2007; Luong et al., 2014) to neural models (Martins et al., 2016, 2017).

QuEst++ (Specia et al., 2015) is a well-known QE pipeline that supports word-level, sentence-level, and document-level QE. Its effectiveness and flexibility make it an attractive candidate for our proposed task. There are two main modules to QuEst++: a feature extractor and a learning module. The feature extractor produces an intermediate representation of the source and translation in a continuous feature vector. The goal of the learning module, given a source and translation pair, is to predict the quality of the translation, either as a label or as a continuous value. This module is trained on example translations that have an assigned score (such as BLEU) and then predicts the score of a new example. QuEst++ offers a range of learning algorithms but defaults to Support Vector Regression for sentence-level QE.

3 Quality Estimation for Interpretation

The default, out-of-the-box, sentence-level feature set for QuEst++ includes seventeen features such as number of tokens in source/target utterances, average token length, n-gram frequency, etc. (Specia et al., 2015). While this feature set is effective for evaluation of MT output, SI output is inherently different—full of pauses, hesitations, paraphrases, re-orderings and repetitions. In the following sections, we describe our methods to adapt QE to handle these phenomena.

3.1 Interpretation-specific Features

To adapt QE to interpreter output, we augment the baseline feature set with four additional types of features that may indicate a struggling interpreter.

Ratio of pauses/hesitations/incomplete words: Sridhar et al. (2013) propose that interpreters regularly use pauses to gain more time to think and as a cognitive strategy to manage memory constraints. An increased number of hesitations or incomplete words in interpreter output might indicate that an interpreter is struggling to produce accurate output. In our particular case, both corpora we use in experiments are annotated for pauses and partial renditions of words.

Ratio of non-specific words: Interpreters often compress output by replacing or omitting common nouns to avoid specific terminology (Sridhar et al., 2013), either to prevent redundancy or to ease cognitive load. For example: "The chairman explained the proposal to the delegates" might be rendered in a target language as "he explained it to them." To capture this, we include a feature that checks for words from a pre-determined seed list of pronouns and demonstrative adjectives.

Ratio of 'quasi-'cognates: In related language pairs, often words of a similar root are orthographically similar, for example "artificial"(EN), "artificiel"(FR) and "artificiale"(IT). Likewise in Japanese, words adapted from English are transcribed in katakana script to indicate their foreign

origin. Transliterated words in interpreted speech could represent facilitated translation by language proximity, or an attempt to produce an approximation of a word that the interpreter did not know. We include a feature that counts the number of words that share at least 50% identical orthography (for EN, FR, IT) or are rendered in the interpreter transcript in katakana (JA).

Ratio of number of words: We further include three features from QuEst++ that compare source and target length and the amount of transcribed punctuation. Information about utterance length makes sense in an interpreting scenario, given the aforementioned strategies of omission and compression of information. A list, for example, may be compressed to avoid redundancy or may be an erroneous omission (Barik, 1994).

3.2 Evaluation Metric

Novice interpreters are assessed for accuracy on the number of omissions, additions and the inaccurate renditions of lexical items and longer phrases (Altman, 1994), but recovery of content and correct terminology are highly valued. While no large corpus exists that has been manually annotated with these measures, they align with the phenomena that MT evaluation tries to solve. One important design decision is which evaluation metric to target in our QE system. There is an abundance of evaluation metrics available for MT including WER (Su et al.), BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and METEOR (Denkowski and Lavie, 2014), all of which compare the similarity between reference translations and translations. Interpreter output is fundamentally different from any reference that we may use in evaluation because interpreters employ a range of economizing strategies such as segmentation, omission, generalization, and reformulation (Riccardi, 2005). As such, measuring interpretation quality by some metrics employed in MT such as BLEU can result in artificially low scores (Shimizu et al., 2013). To mitigate this, we use METEOR, a more sophisticated MT evaluation metric that considers paraphrases and content-function word distinctions, and thus should be better equipped to deal with the disparity between MT and SI. Better handling of these divergences for evaluation of interpreter output, or fine-grained evaluation based on measures from interpretation studies is an interesting direction for future work.

4 Data: Interpretation Corpora

For our EN-JA language data we train the pipeline on combined data from seven TED Talks taken from the NAIST TED SI corpus (Shimizu et al., 2013). This corpus provides human transcribed SI output from three interpreters of low, intermediate and high levels of proficiency denoted B-rank, A-rank and S-rank respectively, with 559 utterances from each interpreter. The corpus also provides written translations of the source speech, which we use as reference translations when evaluating interpreter output using METEOR.

Our EN-FR and EN-IT data are drawn from the EPTIC corpus (Bernardini et al., 2016), which provides source and interpreter transcripts for speeches from the European Parliament (manually transcribed to include vocal expressions), as well as translations of transcripts of the source speech. The EN-FR and EN-IT datasets contain 739 and 731 utterances respectively. While the EPTIC translations are accurate, they were created from an official transcript that differs significantly in register from the source speech. As a proxy for our experiments, we generated translations of the original speech using Google Translate, which resulted in much more qualitatively reliable METEOR scores than the EPTIC translations.

5 Interpreter Quality Experiments

To evaluate the quality of our QE system, we use the Pearson's r correlation between the predicted and true METEOR for each language pair (Graham, 2015). As a baseline, we train QuEst++ on the out-of-the-box feature set (Section 2).

We use k-fold cross-validation individually on EN-JA, EN-FR, and EN-IT source-interpreter language pairs with a held-out development set and test set for each fold. For each experiment setting, we run the experiment for each fold (ten iterations for each set) and evaluate average Pearson's r correlation on the development set.

In our baseline, we extract features based on the default QuEst++ sentence-level features (*baseline*). We ablate baseline features via cross-validation and remove relating bigram, trigram, and punctuation frequency features in the source utterance, creating a more effective trimmed model (*trimmed*).

Subsequently, we add our interpreter features (Section 3.1) and arrive at our *proposed* model. We then repeat each experiment using the test set

	baseline	trimmed	proposed
EN-JA(B-rank)	0.514	0.542	**0.593**
EN-JA(A-rank)	0.487	0.554	**0.591**
EN-JA(S-rank)	0.325	0.334	**0.411**
EN-FR	0.631	0.610	**0.691**
EN-IT	0.569	0.543	**0.576**

Table 1: Pearson's r scores for predicted ME-TEOR for baseline, trimmed and proposed feature sets on the test set (highest accuracy for each dataset indicated in bold).

data from each fold and compare the resulting average Pearson's r scores.

5.1 Results

Table 1 shows our primary results comparing the baseline, trimmed, and proposed feature sets. Our first observation is that, even with the baseline feature set, QE obtains respectable correlation scores, proving feasible as a method to predict interpreter performance. Our trimmed feature set performs moderately better than the baseline for Japanese, and slightly under-performs for French and Italian. However, our proposed, interpreter-focused model out-performs in all language settings with notable gains in particular for EN-JA(A-Rank) (+0.104), achieving its highest accuracy on the EN-FR dataset. Over all datasets, the gain of the proposed model is statistically significant at $p < 0.05$ by the pairwise bootstrap (Koehn, 2004).

5.2 Analysis

We further present two analyses: ablation on the full feature set and a qualitative comparison. Table 2 iteratively reduces the feature set by first removing the 'quasi-'cognate feature (w/o cog); specific words (w/o spec); pauses, hesitations, and incomplete words (w/o fill); and finally sentence length and punctuation differences (w/o length).

Relative difference in utterance length appears to aid Japanese and French above other languages. Cognates are particularly useful in EN-FR and EN-IT; this may be indicative of the corpus domain (European Parliament proceedings being rich in Latinate legalese) or of cognate frequency in those languages. In Japanese, cognates were more indicative of quality for the more skilled interpreter (S-rank). While pauses and hesitations seem to aid the model in EN-FR and EN-IT, they appear to hinder EN-JA.

	w/o cog	w/o spec	w/o fill	w/o len
EN-JA(B)	+0.007	+0.012	+0.016	-0.053
EN-JA(A)	-0.006	-0.011	-0.012	-0.031
EN-JA(S)	-0.014	+0.001	+0.004	-0.061
EN-FR	-0.013	-0.006	+0.007	-0.054
EN-IT	-0.020	+0.002	+0.020	+0.005
Average	-0.009	-0.001	+0.007	-0.039

Table 2: Relative difference in Pearson's r scores for ablated features after removing cognates, specifics, fillers and length difference (cumulative ablation, left to right). Omission and addition are key features distinguishing SI from translation.

Below is a qualitative EN-IT example with METEOR score 0.079 (substantially lower than the average METEOR score across all datasets; 0.262). The baseline predicts a score of 0.127, while our model predicts 0.066:

SOURCE: "*Will the Parliament grant President Dilma Rousseff, on the very first occasion after her groundbaking groundbreaking election and for no sound formal reason, the kind of debate that we usually reserve for people like Mugabe? So, I ask you to remove Brazil from the agenda of the urgencies.*"

INTERP: "*Ehm il Parlamento... dopo le elezioni... daremdar spazio a un dibattito sul ehm sul caso per esempio del presidente Mugabe invece di mettere il Brasile all'ordine del giorno?*"

GLOSS: "*Ehm the Parliament... after the elections... we'll gi- will give way to a debate on the ehm on the case for example of President Mugabe instead of putting Brazil on the agenda?*"

Our model can better capture the issues in this example because it has many interpretation specific qualities (pauses, compression, and omission). This is an example in which a CAI system might offer assistance to an interpreter struggling to produce an accurate rendition.

6 Conclusion

We introduce a novel and effective application of QE to evaluate interpreter output, which could be immediately applied to allow CAI systems to selectively offer assistance to struggling interpreters. This work uses METEOR to evaluate interpreter output, but creation of fine-grained measures to evaluate various aspects of interpreter performance is an interesting avenue for future work.

References

Janet Altman. 1994. *Error Analysis in the Teaching of Simultaneous Interpreting: A Pilot Study*. John Benjamins Publishing Company, Edinburgh, Scotland.

Henry Barik. 1994. *A Description of Various Types of Omissions, Additions and Errors of Translation Encountered in Simultaneous Interpretation*. John Benjamins Publishing Company, North Carolina, USA.

Silvia Bernardini, Adriano Ferraresi, and Maja Milievi. 2016. From EPIC to EPTIC exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies*, 28(1):61–86.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proc. WMT*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proc. WMT*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. HLT*, pages 138–145.

Claudio Fantinouli. 2016. Interpretbank. redefining computer-assisted interpreting tools. In *Proceedings of the 38th Conference Translating and the Computer*, pages 42–55.

Claudio Fantinouli. 2018. Trends in e-tools and resources for translators and interpreters. In *Isabel Durn and Gloria Corpas (eds.)*, pages 153–174.

Yvette Graham. 2015. Improving evaluation of machine translation quality estimation. In *Proc. ACL*, pages 1804–1813.

He He, Jordan Boyd-Graber, and Hal Daumé III. 2016. Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. In *Proc. NAACL*, pages 971–976.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proc. EMNLP*, pages 388–395.

Ngoc-Quang Luong, Laurent Besacier, and Benjamin Lecouteux. 2014. LIG system for word level QE task at WMT14. pages 335–341.

André Martins, Ramon Astudillo, Chris Hokamp, and Fabio Kepler. 2016. Unbabel's participation in the WMT16 word-level translation quality estimation shared task. pages 806–811.

André F.T. Martins, Marcin Junczys-Dowmunt, Fabio N. Kepler, Ramn Astudillo, Chris Hokamp, and Roman Grundkiewicz. 2017. Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218.

B Moser-Mercer, A Knzli, and M Korac. 1998. Prolonged turns in interpreting: Effects on quality, physiological and psychological stress. In *INTERPRETING*, pages 47–65.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

Alessandra Riccardi. 2005. On the evolution of interpreting strategies in simultaneous interpreting. *Meta*, 50(2):753–767.

Hiroaki Shimizu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2013. Constructing a speech translation system using simultaneous interpretation data. In *Proc. IWSLT*, pages 212–218.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *Proc. ACL*, pages 115–120.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

Vivek Kumar Rangarajan Sridhar, John Chen, Srinivas Bangalore, Andrej Ljolje, and Rathinavelu Chengalvarayan. 2013. Segmentation strategies for streaming speech translation. In *Proc. NAACL*, pages 230–238.

Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. A new quantitative quality measure for machine translation systems. In *Proceedings of the 14th Conference on Computational Linguistics*.

Nicola Ueffing and Hermann Ney. 2007. Word-level confidence estimation for machine translation. *Computational Linguistics*, 33(1):9–40.