PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 121 JUNE 2024 5–14

Revamping the SLTev Tool for Evaluation of Spoken Language Translation

Michelle Elizabeth, Ondřej Bojar

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

This article describes recent improvements of SLTev, a tool for automatic evaluation of machine translation, speech recognition and speech translation systems. The changes include the implementation of the COMET score for evaluation of machine translation and spoken language translation outputs as well as a fix for the problematic delay calculation for repeated words which favoured longer segments. Additionally, the system outputs of the IWSLT 2022 shared task have also been evaluated using SLTev and a comparison study with another speech evaluation toolkit, SimulEval, has been done.

1. Introduction

Spoken Language Translation or SLT is a prominent task in NLP. In the so-called cascaded approach, it involves translation of speech into various languages and combines automatic speech recognition in the source language and then machine translation into the target languages. On the end-to-end approaches, the intermediate transcription in the source language is not explicitly considered. As with any application in NLP, it is necessary to evaluate the results produced with suitable metrics.

SLTev[1] or Spoken Language Translation Evaluation tool (Ansari et al., 2021) performs the evaluation of the outputs of spoken language translations by reporting the quality, latency, and stability of a candidate output based on its time-stamped transcript and reference translation into a target language.

[1] https://github.com/ELITR/SLTev

Corresponding author: bojar@ufal.mff.cuni.cz
Cite as: Michelle Elizabeth, Ondřej Bojar. Revamping the SLTev Tool for Evaluation of Spoken Language Translation. The Prague Bulletin of Mathematical Linguistics No. 121, 2024, pp. 5–14. doi: 10.14712/00326585.028.

This article gives an overview of the work done to improve the SLTev tool. Our main contributions include the addition of a new metric, COMET for evaluating machine translation and the fix for delay calculation so that it takes into account the timestamps of repeated tokens. Also, an evaluation of the system outputs of IWSLT 2022 (Anastasopoulos et al., 2022) was done using the SLTev tool, comparing the scores also to the results of SimulEval, the tool used officially for IWSLT 2022.

We first look at the SLTev tool and the metrics it uses for evaluation. Then we discuss the issues and challenges existing in the current implementation of the tool and the work done to mitigate these issues and improve the tool. Finally, we evaluate the IWSLT 2022 outputs using SLTev.

2. Background

The SLTev tool is an open source tool for evaluating SLT outputs against reference translations and time-stamped source transcripts. It was developed as part of the European Live Translator (ELITR) project (Franceschini et al., 2020) and provided three metrics namely SacreBLEU (Post, 2018) for measuring quality, Flicker for measuring the stability and Delay for measuring the latency of SLT outputs.

Translation quality is estimated using the SacreBLEU tool applied in three different ways within SLTev. The first one considers all completed segments as a single joint segment and compares it with the reference which is also considered as a single concatenated segment. The second variant uses mwerSegmenter (Matusov et al., 2005) to compare candidate and segmented reference outputs. The final variant relies on time-span quality and divides the whole document into chunks or segments of a fixed duration which are then separately evaluated using BLEU and also averaged for the score of the whole document.

Flicker assesses the amount of intermediate output updates which can distract the user by counting the number of words after the first difference between two consecutive output updates. Flicker is reported in two variants: average revision count per second and normalised revision count which are described by Ansari et al. (2021).

The final measure is delay which measures the difference between the time that a target word was displayed and an estimate of when it should have been displayed given the source transcript. The delay is calculated using two approaches. The first one is proportional which estimates the timing of each source word based on partial segments in the golden transcript. These times are then passed to the words in the reference translation proportionally along the sequence of words. The second approach uses automatic word alignment between the source and reference translations to account for word order differences across languages.

2.1. SLTev vs. SimulEval

Another toolkit that is similar to SLTev and has been developed for evaluating simultaneous translation is SimulEval (Ma et al., 2020). It has been used as the evaluation toolkit for the IWSLT Shared Task since its first edition in 2020.

SimulEval is based on a client-server scheme in which the server sends the source input when requested by the client, receives the translation for evaluation from the client and reports various metrics pertaining to translation quality and latency. The client is composed of an agent and a state. The agent is responsible for executing the system's policy and the state tracks the necessary information for executing the policy when generating the translation. SLTev on the other hand uses a time-stamped golden transcript in the source language, a reference translation and candidate output in the target language to evaluate the translation quality, latency and stability off-line, i.e. from logs and without running the system again.

SimulEval reports BLEU, TER and METEOR for evaluating translation quality and has adapted Average Proportion, Average Lagging and Differentiable Average Lagging for speech translation. However, it does not support any assessment of output; the evaluated systems are not permitted to their older outputs in any way. SLTev reports stability using the flicker metric along with measuring translation quality and latency using SacreBLEU and delay respectively as described previously. An evaluation of the IWSLT 2022 using SLTev and a comparison of the results with the official SimulEval results is reported later in Section 3.3.

3. Improvements to the Current Implementation of SLTev

This section describes some of the issues that the current implementation of SLTev had and the work done to improve the tool.

3.1. Delay Computation

In the existing implementation of delay computation, there was an issue in how the time-stamps were assigned to repeated words in a segment. This problem has been reported by Amrhein and Haddow (2022). The following example can be used to explain the problem:
P 13.18 O
P 14.18 O horror,
P 15.18 O horror, terror, horror
C 16.18 O horror, horror, horror.

SLTev assigned the time stamp of 14.18 to all occurrences of the word "horror", i.e. it assigned the token the time-stamp of its first occurrence even though later updates actually discarded some of these occurrences. When translating longer segments easily consisting of multiple sentences, the likelihood of encountering tokens that were

previously seen increases. In such cases, all of these tokens would be assigned the time-stamp of their first occurrence. Hence, this tends to favour longer segments in the translation process.

3.2. COMET

COMET (Rei et al., 2020), which stands for Cross-lingual Optimized Metric for Evaluation of Translation, is a popular neural framework for training multilingual machine translation evaluation models. Typically, COMET models are trained with the objective of predicting quality scores for translations. These scores are usually normalized through a z-score transformation and serve as a valuable metric for ranking translations and systems based on their quality.

The COMET library has several evaluation models and we use the default model Unbabel/wmt22-comet-da (Rei et al., 2022). This model utilizes a reference-based regression methodology and is constructed using the XLM-R framework. It has undergone training on direct assessments from WMT17 to WMT20, offering scores within the 0 to 1 range. A score of 1 indicates a perfect translation.

The Unbabel/wmt22-comet-da model is available on HuggingFace and can be downloaded. A list of dictionaries containing the source, candidate translation and the reference is given as input to the model. It generates scores for each set of source, candidate, reference triplet and also reports an overall system score.

For SLTev, the segmented candidate sentences were concatenated together to form one single segment. The same was done for the source and reference segments in order to generate an overall score for the document. The generated system score which is reported by the model in the range $[0, 1]$ has been scaled to $[0, 100]$ in order to be consistent with the SacreBLEU reporting in SLTev.

One issue that was observed during the implementation was that internet connection was necessary in order to download the model to the local system. Currently, this situation is being handled in a way that does not disrupt the flow of the evaluation by handling the exception where the download has failed and moving on to the next metrics.

3.3. Evaluation of the IWSLT 2022 System Outputs

The SLTev tool was used to run an evaluation of the outputs by the models submitted to the IWSLT 2022 Simultaneous Speech Translation task. The language pair was English to German and the outputs of five systems namely CUNI-KIT (Polák et al., 2022), FBK (Gaido et al., 2022), HW-TSC (Wang et al., 2022), NAIST (Fukuda et al., 2022) and UPV (Iranzo-Sánchez et al., 2022) were evaluated. Each system has produced outputs for three latency regimes — high, medium and low — determined by a maximum latency threshold measured by Average Lagging on the Must-C tst-COMMON set.

Model	Delay Without Partials	Delay With Partials	SacreBLEU	COMET	Flicker
CUNI-KIT.high	61.24	62.49	**25.59**	**66.57**	0.00
CUNI-KIT.medium	71.39	71.84	23.32	64.98	0.00
CUNI-KIT.low	79.72	79.93	17.76	60.01	0.00
FBK.high	90.25	91.91	18.57	54.12	0.00
FBK.medium	58.51	60.20	15.85	51.22	0.00
FBK.low	51.07	52.23	8.62	41.18	0.00
HW-TSC.high	50.31	50.91	9.88	35.09	0.00
HW-TSC.medium	60.24	60.85	9.82	35.3	0.00
HW-TSC.low	65.44	66.1	8.31	33.94	0.01
NAIST.high	77.82	79.37	9.00	38.33	0.00
NAIST.medium	**27.79**	**28.73**	9.16	38.00	0.00
NAIST.low	38.15	38.34	7.03	39.75	0.00
UPV.high	238.31	241.12	22.88	62.29	0.00
UPV.medium	148.25	150.31	19.49	59.97	0.00
UPV.low	176.52	179.75	12.81	52.06	0.00

Table 1. SLTev Evaluation of IWSLT 2022 System Outputs

Latency	High	Medium	Low
Pearson Correlation	0.89	0.864	0.891

Table 2. Pearson Correlation of IWSLT 2022 System Outputs with respect to SimulEval and SLTev Results

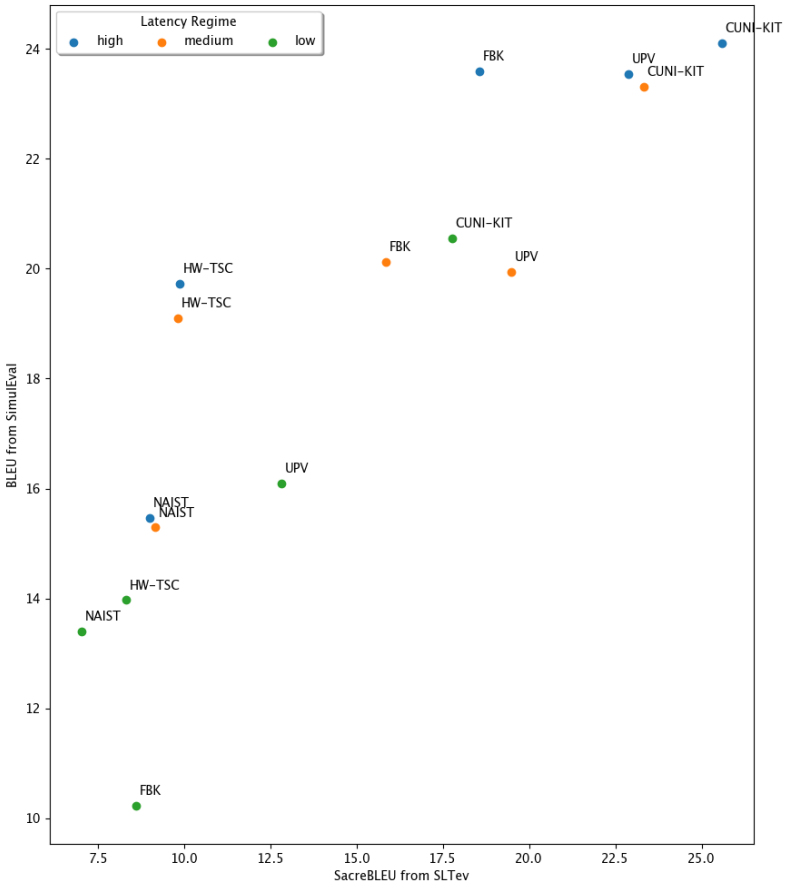

Figure 1. Scatterplot showing the BLEU scores reported by SLTev and SimulEval for IWSLT 2022 systems for three latency regimes

IWSLT 2022 used the SimulEval toolkit (Ma et al., 2020) for evaluating the quality and latency of the submissions. The metrics they used were BLEU for measuring translation quality and average proportion (AP), average lagging (AL) and differentiable average lagging (DAL) for measuring translation latency. Using SLTev, the delay measurements calculated using the partial segments and the one which considers only the completed segments, SacreBLEU, COMET and Flicker have been measured. We can see that the translation quality is highest in each system for the high latency regime except for NAIST which has the best BLEU score for medium latency regime. The CUNI-KIT system in the high latency regime has the best translation quality in terms of both BLEU and COMET. NAIST (medium) has the least average delay whereas UPV systems have the highest delays among all the systems. The metrics are reported in Table 1.

The BLEU scores reported by SimulEval and SLTev in the three latency regimes — high, medium and low — have also been reported as a scatter plot in Figure 1. We can observe that the general trend is that SLTev has scored the systems lower than the scores reported by SimulEval except for CUNI-KIT in the high latency regime. HW-TSC and NAIST have been scored much lower by SLTev than by SimulEval, a difference in the range of approximately 5–10 points. This can also be seen for FBK though not to the extent of HW-TSC and NAIST. CUNI-KIT and UPV have similar scores reported by SLTev and SimulEval. Table 2 reports the Pearson correlation for scores of the systems in the three latency regimes.

4. Conclusion and Future Work

SLTev is a comprehensive tool for evaluating the quality of spoken language translation. We wish it became the standard toolkit with a wide adoption.

The work done reported in this article is just the beginning, there is more room for improvement. The implementation of COMET score can be enhanced further by reporting segment-level scores as well. It would also be beneficial to figure out how to download the COMET model available in HuggingFace locally when installing SLTev and not having to rely on a stable Internet connection to generate the score. The bug fix for delay computation should give more accurate results and will no longer favour longer segments since the time-stamps of repeated tokens are being accurately calculated. Additional metrics relevant to translation can be added including average lagging and chrF3 (Popović, 2015).

The tool can be made more versatile by making it platform independent. Right now, it relies on mwerSegmenter which can only be run on Linux systems. Word-error-rate-based segmentation is thus not preformed for quality evaluation when used on other platforms. The readability and reusability of the code can be improved by using more Pythonic constructions. Also, writing and maintaining unit tests, implementing a proper error handling module and detailed logging are some other ways in which the tool can be made more user and developer friendly.

Bibliography

Amrhein, Chantal and Barry Haddow. Don't Discard Fixed-Window Audio Segmentation in Speech-to-Text Translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 203–219, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.13.

Anastasopoulos, Antonios, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. Findings of the IWSLT 2022 Evaluation Campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.10. URL https://aclanthology.org/2022.iwslt-1.10.

Ansari, Ebrahim, Ondřej Bojar, Barry Haddow, and Mohammad Mahmoudi. SLTEV: Comprehensive Evaluation of Spoken Language Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 71–79, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.9. URL https://aclanthology.org/2021.eacl-demos.9.

Franceschini, Dario, Chiara Canton, Ivan Simonini, Armin Schweinfurth, Adelheid Glott, Sebastian Stüker, Thai-Son Nguyen, Felix Schneider, Thanh-Le Ha, Alex Waibel, Barry Haddow, Philip Williams, Rico Sennrich, Ondřej Bojar, Sangeet Sagar, Dominik Macháček, and Otakar Smrž. Removing European Language Barriers with Innovative Machine Translation Technology. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 44–49, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-64-1. URL https://aclanthology.org/2020.iwltp-1.7.

Fukuda, Ryo, Yuka Ko, Yasumasa Kano, Kosuke Doi, Hirotaka Tokuyama, Sakriani Sakti, Katsuhito Sudoh, and Satoshi Nakamura. NAIST Simultaneous Speech-to-Text Translation System for IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 286–292, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.25. URL https://aclanthology.org/2022.iwslt-1.25.

Gaido, Marco, Sara Papi, Dennis Fucci, Giuseppe Fiameni, Matteo Negri, and Marco Turchi. Efficient yet Competitive Speech Translation: FBK@IWSLT2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 177–189, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.13. URL https://aclanthology.org/2022.iwslt-1.13.

Iranzo-Sánchez, Javier, Javier Jorge Cano, Alejandro Pérez-González-de Martos, Adrián Giménez Pastor, Gonçal Garcés Díaz-Munío, Pau Baquero-Arnal, Joan Albert Silvestre-Cerdà, Jorge Civera Saiz, Albert Sanchis, and Alfons Juan. MLLP-VRAIN UPV systems for

the IWSLT 2022 Simultaneous Speech Translation and Speech-to-Speech Translation tasks. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 255–264, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.22. URL https://aclanthology.org/2022.iwslt-1.22.

Ma, Xutai, Mohammad Javad Dousti, Changhan Wang, Jiatao Gu, and Juan Pino. Simuleval: An evaluation toolkit for simultaneous translation. In *Proceedings of the EMNLP*, 2020. doi: 10.18653/v1/2020.emnlp-demos.19.

Matusov, Evgeny, Gregor Leusch, Oliver Bender, and Hermann Ney. Evaluating Machine Translation Output with Automatic Sentence Segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA, October 24-25 2005. URL https://aclanthology.org/2005.iwslt-1.19.

Polák, Peter, Ngoc-Quan Pham, Tuan Nam Nguyen, Danni Liu, Carlos Mullov, Jan Niehues, Ondřej Bojar, and Alexander Waibel. CUNI-KIT System for Simultaneous Speech Translation Task at IWSLT 2022. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 277–285, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.24. URL https://aclanthology.org/2022.iwslt-1.24.

Popović, Maja. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049.

Post, Matt. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://www.aclweb.org/anthology/W18-6319.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 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, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Rei, Ricardo, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52.

Wang, Minghan, Jiaxin Guo, Yinglu Li, Xiaosong Qiao, Yuxia Wang, Zongyao Li, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang, and Ying Qin. The HW-TSC's Simultaneous Speech Translation System for IWSLT 2022 Evaluation. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 247–254, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.21. URL https://aclanthology.org/2022.iwslt-1.21.

Address for correspondence:
Ondřej Bojar
bojar@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics,
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic