Reproducing Speech to Text Translation Results from two prominent foundation models: Whisper and SeamlessM4T

Shrey Jasuja
NYU Tandon School of Engineering
New York, USA
shrey.jasuja@nyu.edu

Fraida Fund
NYU Tandon School of Engineering
New York, USA
ffund@nyu.edu

ABSTRACT

In this study, we aim to reproduce the results of Speech-to-text translations from large multitask models in the papers:"SeamlessM4T: Massively Multilingual & Multimodal Machine Translation"[2] and "Robust Speech Recognition via Large-Scale Weak Supervision"[7]. We restrict ourselves to claims corresponding to translations from other languages to English. During our study, we were able to reproduce most of the claims, but some results weren't well-aligned due to missing reference to the decoding strategy used. While reproducing the results we also try to evaluate the ease of reproducibility.

CCS CONCEPTS

• **Computing methodologies** → **Machine translation**; **Speech recognition**; • **General and reference** → Evaluation.

KEYWORDS

S2TT, multitask, speech systems, machine translation, reproducibility, speech-to-text translation

ACM Reference Format:
Shrey Jasuja and Fraida Fund. 2024. Reproducing Speech to Text Translation Results from two prominent foundation models: Whisper and SeamlessM4T . In *Proceedings of 2024 ACM Conference on Reproducibility and Replicability, (ACM REP '24).* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 METHODOLOGY

We identified certain claims stipulating results for translations from other languages to English. These claims were found to be based on CoVoST2 [9] and FLEURS[4] speech datasets. CoVoST 2 includes data for 21 languages with their translations in English and FLEURS consists of 101 languages supporting translations to English, as well as n-way translations.

We could only evaluate those models in the claim where models were open-sourced because of two core hurdles. Firstly, most of the authors have proprietary datasets that are not publicly available. Thus, even if they discuss the methodology of their study it is not possible to replicate the model architecture to generate exact results. Second, with the increasing complexity of these large multitask models, it is difficult to put in the computing resources required to train these models as intended even if the data is available.

We found that the authors of Whisper[7], SeamlessM4T [2] and XLS-R[1] models have released these models on HuggingFace [10], which allowed us to evaluate the claims.

We used Chameleon[5] which is an open testbed for our computation requirements. Chameleon has a bare-metal reconfiguration system that enables users to control the kernels, manage resources, and console access, allowing us to obtain consistent results. We used RTX 6000 GPU for our study.

Table 1: Comparison of reproduced BLEU scores against the original claim in Seamless-M4T[2]. The values in the parenthesis are the original values in the claim.

Model	size	FLEURS X→eng (n=81)	CoVoST 2 X→eng (n=21)
XLS-R-2B-S2T	2.6B	()	22.1 (22.1)
WHISPER-LARGE-v2	1.5B	16.7 (17.9)	29.2 (29.1)
AUDIOPaLM-2-8B-AST	8.0B	(19.7)	(37.8)
SEAMLESSM4T-MEDIUM	1.2B	20.3 (20.9)	31.3 (29.8)
SEAMLESSM4T-LARGE	2.3B	23.4 (24.0)	34.3 (34.1)

2 RESULTS

The results we reproduced when evaluating the performance of the models on the CoVoST 2 and FLEURS datasets, were consistent with the claims established in the papers. We used the BLEU score from SacreBleu as an evaluation metric, consistent with the metrics used in the claims. We reproduced a total of 4 claims, Table 1 shows one of the claims that we reproduced compared to the original claim. In this table, we have omitted the columns depicting results for eng → X translations as we only investigated X → eng translations. We couldn't reproduce results for AUDIOPaLM as it is a proprietary model. We noticed some slight deviation in the performance of the Whisper model on the FLEURS dataset which we suspect was due to a different decoding strategy used while evaluating, the parameters for which weren't explicitly stated in the paper.

3 EASE OF REPRODUCIBILITY

During the course of this study, there were some parts where we felt at ease while reproducing the results. On the flip side, we also encountered significant challenges, some of which we were able to mitigate, while others were beyond our control.

3.1 The Easy Part

The availability of open-sourced models on HuggingFace as a single platform ensured easy access to the models. They also mention documentation on the model card on how to use the models, example

ACM REP '24, June 18–20, 2024, Rennes, France

ACM ISBN 978-1-4503-XXXX-X/24/06
https://doi.org/XXXXXXX.XXXXXXX

code, and other useful information. This led to fast-paced and less effort development for the inference pipeline.

The authors of all the papers also included detailed appendices, containing granular-level results for each language, which proved to be very useful. In the Seamless paper, we also found the evaluation IDs used for the claim on their GitHub repository for the project, this enabled us to use those while reproducing claims.

3.2 The Challenges

- **Unavailibity of some models and their training dataset**: We found that certain models like Maestro [3], AudioPaLM [8] and others didn't have their checkpoints released. So, we couldn't reproduce their results. Although the authors did a great job documenting model architecture and methodology, coupled with a lack of training dataset, they didn't enable us to exactly reproduce the results.
- **Inconsistent language codes/verbatim**: We required language codes to access datasets and during model inference while specifying source and target language. We noticed that different papers and datasets specify different language codes, leading to non-uniformity. For example, authors of the Whisper model and CoVoST 2 dataset use ISO 639-1 two-letter language code, FLEURS dataset use BCP-47 codes with two-letter primary subtag and region subtag based on a two-letter country code from ISO 3166-1 alpha-2, NLLB model uses BCP-47 codes with two-letter primary subtag and script subtag based on a four-letter script code from ISO 15924 and Seamless model uses ISO 639-3 language codes. We also noticed that different dialects of the same language increase complexity when trying to establish cross-mapping or directly use a language name. There was also an instance in he SeamlessM4T paper where the authors initially listed 'nno' and 'nob' as Norwegian language codes, but later used 'nor' for data statistics. After thorough verification, we later revealed that 'nob' was actually used in the implementation.
- **Computation time**: It is often time-consuming to evaluate these models on large datasets on single node instances in academic settings. With added complexity because of speech as an input modality, the inference took a lot of time for use, consuming 1.5 days on average per model per dataset. During research at private companies with unlimited resources, they tend to use GPU clusters for such tasks, speeding up the inference. Since we wanted to make our work reproducible we chose single nodes, accessible to everyone to build upon our study.

4 ARTIFACTS USED

We used the test split of the following publicly available datasets for $X \rightarrow$ eng translations when reproducing the results:

- CoVoST 2: We used the multilingual voices from the Common Voice dataset, Common Voice Corpus 4, and its annotations from the CoVoST 2[9] dataset at HuggingFace datasets.
- FLEURS: We used the FLEURS[4] dataset implementation on HuggingFace datasets. We matched the voices to appropriate IDs in English as it supported N-way translations using transcriptions.

We obtained pre-trained models released by individual authors on HuggingFace:

- SeamlessM4T: SeamlessM4T[2] models, including the model weights for SeamlessM4T-Large (2.3B) and SeamlessM4T-Medium (1.2B) and its inference code from HuggingFace.
- Whisper: Whisper[7] Large-v2 (1.55B) model and inference code from Whisper implementation on Whisper's GitHub repository.
- XLS-R: XLS-R[1] model fine-tuned on CoVoST2 dataset. We used model weights and inference code from 'wav2vec2-xls-r-2b-21-to-en' model card.
- NLLB[6]: The 1.3B model to set up a cascaded S2TT pipeline with Whisper Large-v2 to compare its outcomes with direct translation as cited in the claims.

5 ARTIFACTS CREATED

In an attempt to make our study easily reproducible and allow users to build upon it, we have released our source code on GitHub at *https://github.com/shreyjasuja/re_s2tt*. The source code provides Jupyter notebooks to initialize the environment on Chameleon, install required dependencies, download and manage datasets, run inference, and analyze and compare final results to original claims. We also published it as a Trovi artifact on Chameleon to ease access to the project on Chameleon testbeds.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2226408.

REFERENCES

[1] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296* (2021).

[2] Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. SeamlessM4T-Massively Multilingual & Multimodal Machine Translation. *arXiv preprint arXiv:2308.11596* (2023).

[3] Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro Moreno, Ankur Bapna, and Heiga Zen. 2022. Maestro: Matched speech text representations through modality matching. *arXiv preprint arXiv:2204.03409* (2022).

[4] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 798–805.

[5] Kate Keahey, Jason Anderson, Zhuo Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mert Cevik, Jacob Colleran, Haryadi S. Gunawi, Cody Hammock, Joe Mambretti, Alexander Barnes, François Halbach, Alex Rocha, and Joe Stubbs. 2020. Lessons Learned from the Chameleon Testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association.

[6] James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam et al. NLLB Team, Marta R. Costa-jussà. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. (2022).

[7] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[8] Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925* (2023).

[9] Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310* (2020).

[10] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).