Table of Contents

Abstract

Multilingual Speech-to-Text Translation (ST) plays a pivotal role in bridging linguistic barriers by converting spoken language into written text across different languages. This project aims to develop a robust ST model tailored for low-resource Indian languages, specifically targeting the Indo-Aryan and Dravidian language families. The dataset used will consist of speeches from conferences and TED Talks, along with their corresponding transcriptions in English (source language) and translations in Hindi, Bengali, and Tamil (target languages). By addressing the scarcity of data and attention imbalances in low-resource languages, the project strives to create an efficient ST system capable of real-world deployment. Additionally, leveraging creative approaches, such as utilizing existing resources in related languages or word-level translation resources, will be explored to enhance translation accuracy.

Important Keywords: Multilingual; Speech-to-Text Translation; Low-resource languages; Indian languages; Indo-Aryan; Dravidian.

1. Introduction

Multilingual Speech-to-Text Translation (ST) is indispensable for facilitating communication across diverse linguistic contexts. While recent advancements have shown remarkable progress, many dialects and low-resource languages still lack sufficient parallel data for effective supervised learning. Creative approaches are essential to overcome this challenge, such as leveraging resources from related languages or utilizing word-level translation resources and raw audio. This project aims to address these gaps by developing an End-to-End (E2E) or Cascaded ST model for low-resource Indian languages, including Hindi, Bengali, and Tamil.

2. Motivation

The scarcity of translators proficient in multiple languages, especially in low-resource settings, highlights the urgent need for multilingual speech translation (ST) systems. In regions like India, characterized by a multitude of languages, the development of dedicated models for Indian languages is essential for effective communication. This project aims to advance speech translation technology for a wide range of languages. Our ultimate goal is to foster inclusivity and accessibility through the creation of robust ST models.

This project is fueled by a strong commitment to address significant challenges in speech translation, with a particular focus on languages spoken in India. In today's interconnected world, the ability to communicate across different languages is crucial. However, the shortage of translators who can handle multiple languages, especially in resource-constrained areas, presents a major obstacle. This year, the 21st International Conference on Spoken Language Translation (IWSLT) has released 8 shared tasks. This project is one of those tracks, named the Indic track, and we are excited to embrace this challenge. Our aim is to ensure that everyone, regardless of the language they speak, can communicate effectively. The Indic track at the conference, which encompasses the diverse languages spoken in India, is of particular interest to us. We are deeply passionate about leveraging technology to promote inclusivity and accessibility worldwide. By actively participating in IWSLT 2024, we hope to develop improved tools for speech translation, with a specific focus on Indian languages. Our goal is to empower individuals who have previously lacked access to such technology. [1]

While we acknowledge the significant challenges ahead, such as the shortage of multilingual individuals and insufficient data for certain languages, we are determined to find innovative solutions. Collaboration with others in the field will be key to overcoming these obstacles and making meaningful progress. Participating in IWSLT 2024 provides us with an invaluable opportunity not only to advance our project but also to connect with others, exchange ideas, and inspire further advancements in this field.

In summary, our project is dedicated to using technology as a tool to break down language barriers and foster greater connectivity among people. With perseverance, collaboration, and a heartfelt commitment, we are confident that we can make a tangible difference in facilitating communication across diverse languages.

3. Related Work

Prior research in ST has primarily focused on high-resource languages, leaving many dialects and low-resource languages underserved. The lack of parallel data poses a significant challenge in training supervised learning models for these languages. However, recent efforts have demonstrated the effectiveness of leveraging existing resources from related languages and employing innovative approaches to enhance translation accuracy. [1]

The 20th International Conference on Spoken Language Translation (IWSLT) organized shared tasks targeting nine scientific challenges in spoken language translation (SLT). These tasks covered a wide spectrum, including simultaneous and offline translation, automatic subtitling and dubbing, speech-to-speech translation, multilingual translation, dialect and low-resource speech translation, and formality control. The conference witnessed substantial interest with a total of 38 submissions from 31 teams, evenly distributed between academia and industry. [2]

The focal point of the 2023 IWSLT Evaluation Campaign was offline SLT, entailing the translation of audio speech from one language to text in another without time constraints. It comprised three sub-tasks for translating English into German, Japanese, and Chinese. Participants were granted the flexibility to employ either cascade architectures, amalgamating automatic speech recognition (ASR) and machine translation (MT) systems, or end-to-end approaches directly translating input speech. [2]

Principal objectives were twofold: firstly, to gauge the performance disparity between cascade and end-to-end systems, and secondly, to evaluate SLT technology's competence in handling intricate scenarios like spontaneous speech, noisy audio conditions, and overlapping speakers. The introduction of new test sets, encompassing ACL presentations and press conferences/interviews, aimed at a comprehensive assessment of system efficacy. [2]

Training data conditions spanned from constrained to unconstrained, offering varying levels of access to training resources. Development data encompassed TED talks, ACL presentations, and interviews from the European Parliament Multimedia Centre. System evaluations were conducted employing BLEU and COMET metrics, supplemented by human evaluation on the best-performing submissions. [3]

Ten teams partook in the offline task, collectively submitting 37 runs. A plethora of techniques were employed across these submissions, including cascade and direct models, leveraging large language models, multimodal representations, data augmentation, ensemble methods, and advanced training strategies. Evaluation criteria emphasized the attainment of high translation quality across diverse language pairs and challenging scenarios. [2]

4. Proposed Work

4.1 Conceptual Design Diagram

The conceptual design of our application revolves around the seamless integration of audio transcription and translation capabilities, aimed at providing users with an efficient and user-friendly experience. The diagram below illustrates the core components and flow of our application's functionality

4.2 Key Components of the app are

Input Module

- Responsible for receiving audio files
- Validates and preprocesses the input data for further processing.

Audio Processor and Transcription Module

- Responsible to cleaning the audio file
- Uses ResembleAI for Noise reduction, Restoring distortion and enhancing speech bandwidth
- Uses OpenAI's Whisper model for accurate transcription of audio files.

Translation Module

- Integrates the Helsinki model for achieving multilingual translation of the transcribed text, including Hindi and Tamil.
- Fine Tuning of pre trained translator model to enhance the result quality. [4]

Output Module

- Performs syntax correction and eliminates any detectable hallucination by the model.
- Delivers the translated text to users in their desired format, such as text files.

4.3 Workflow

The workflow of the application begins with the Input Processing stage, where users upload audio files or provide input through supported channels. This initial step serves as the gateway for input data, ensuring its integrity and validity before proceeding further. The Input Module undertakes the crucial task of verifying and preprocessing the audio data, preparing it for subsequent processing stages by addressing any potential inconsistencies or errors.

Following input processing, the Transcription and Translation stages come into play, enabling the transformation of audio content into translated text. The Transcription Module utilizes advanced algorithms to convert audio files into text, maintaining a high level of accuracy and reliability throughout the process. Meanwhile, the Translation Module leverages the Helsinki model to translate the transcribed text into specific languages, ensuring linguistic precision and preserving contextual nuances to facilitate effective communication across language barriers. [4]

Once the transcription and translation processes are complete, the Output Delivery stage takes over, presenting the translated text to users through the Output Module. This component enables seamless access and utilization of the translated content, offering users the flexibility to download the text or integrate it directly into their workflows for further use. By providing a user-friendly interface and facilitating easy dissemination of translated content, the application empowers users to overcome language barriers and engage in effective cross-cultural communication.

5. Experimental Setup

The SpeechSync system design flow aims to convert audio data into high-quality translated text. Let’s look at the environmental settings

5.1 Environment Settings

- **Prerequisites**
  - Python 3.11
  - ffmpeg (command-line tool)

- **Installing ffmpeg**
  - (Ubuntu) sudo apt update && sudo apt install ffmpeg
  - (MacOS) brew install ffmpeg
  - (Windows) choco install ffmpeg

- **App Installation**

  Step 1 - **Clone the repository**

  Shell - git clone git@github.com:ayushannand/SpeechSync.git

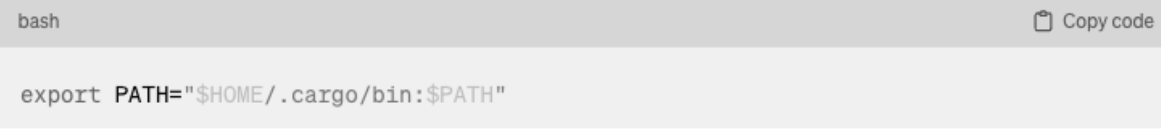
  Step 2 - **Create a virtual environment**

  Shell - python3 -m venv env

  Step 3 - **Activate the virtual environment**

  Shell - source env/bin/activate

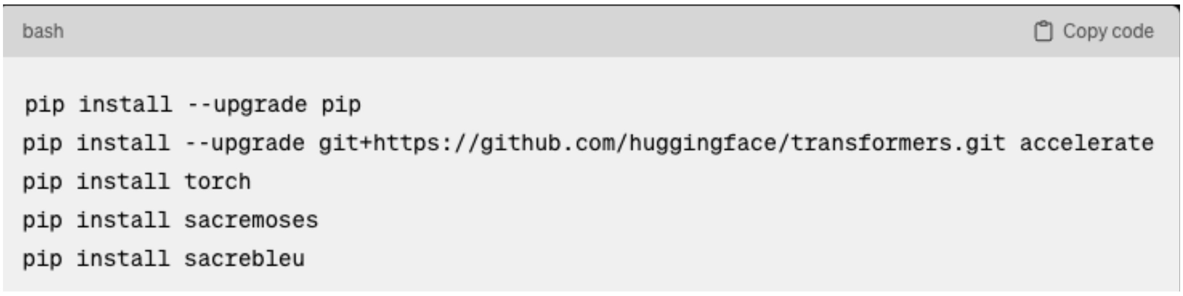
- **Install Rust**



```bash
export PATH="$HOME/.cargo/bin:$PATH"
```

Fig. 3 shows command to install rust

- **Installation of dependencies**



```bash
pip install --upgrade pip
pip install --upgrade git+https://github.com/huggingface/transformers.git accelerate
pip install torch
pip install sacremoses
pip install sacrebleu
```

Fig. 4 shows command to install dependencies

5.2 Metrics :

SacreBLEU is a tool designed for easy computation of BLEU scores, aiming to provide shareable, comparable, and reproducible results. Inspired by Rico Sennrich's multi-bleu-detok.perl, SacreBLEU produces official Workshop on Machine Translation (WMT) scores while working with plain text. It is equipped with knowledge of standard test sets and manages the downloading, processing, and tokenization of data. [3]

How to Use:

To utilize SacreBLEU, provide a set of predictions and references as inputs, along with optional parameters.

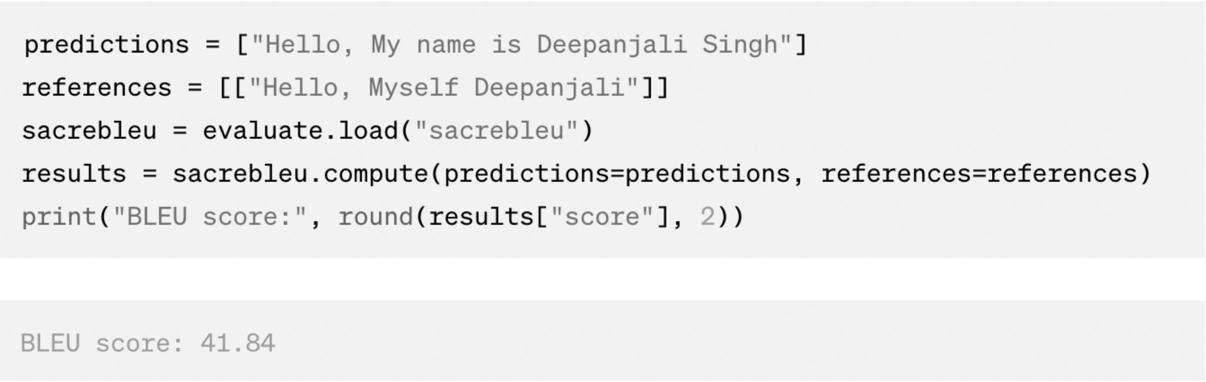
Inputs:

1. predictions (list of str): Predicted translations, each tokenized into a list of tokens.
2. references (list of list of str): References for predictions, organized as a list of lists where each sublist corresponds to a prediction's references.
3. smooth_method (str): Method for smoothing BLEU scores.
4. smooth_value (float): Smoothing value, applicable for certain smoothing methods.
5. tokenize (str): Tokenization method for BLEU computation.
6. lowercase (bool): If True, enables case-insensitivity by lowercasing inputs.
7. force (bool): If True, insists that tokenized input is detokenized.
8. use_effective_order (bool): If True, excludes n-gram orders with zero precision.

Output Values:

1. score: BLEU score
2. counts: Counts
3. totals: Totals
4. precisions: Precisions
5. bp: Brevity penalty
6. sys_len: Length of predictions
7. ref_len: Length of references

Example -



```
predictions = ["Hello, My name is Deepanjali Singh"]
references = [["Hello, Myself Deepanjali"]]
sacrebleu = evaluate.load("sacrebleu")
results = sacrebleu.compute(predictions=predictions, references=references)
print("BLEU score:", round(results["score"], 2))
```

```
BLEU score: 41.84
```

Fig. 5 shows blue score demonstration

6. Current Status and Future Plans of Project Work

6.1 Current Status

- Rigorous experimentation conducted to identify effective models for speech translation.
- Extensive preprocessing of data performed to ensure quality and suitability for training.
- Establishment of a robust pipeline including code development and workflow setup.
- Targeted training and experimentation focused on one language for in-depth analysis.
- Close monitoring of performance metrics and numerical evaluations for model assessment.

6.2 Future Plans

- Expansion of the system to include multiple languages such as Tamil and Bengali.
- Fine-tuning of existing models and experimentation across diverse linguistic domains.
- Implementation of manual evaluation alongside quantitative metrics to ensure translation accuracy.
- Rigorous testing to assess the performance of the expanded system across different language pairs.
- Aim to further enhance capabilities and contribute to the advancement of multilingual speech translation technology.

7. Conclusion

In conclusion, this project is committed to advancing multilingual speech translation (ST) technology, with a specific focus on low-resource Indian languages. Through the creation of dedicated datasets and the development of robust models, our aim is to facilitate seamless communication and accessibility across diverse linguistic communities, ultimately promoting inclusivity and empowerment.

Building upon these objectives, our project seeks to develop specialized ST models tailored to low-resource Indian languages, thereby driving progress in speech translation technology.

References

[1] Elizabeth Salesky, Marcello Federico, and Marine Carpuat. 2023. Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023). Association for Computational Linguistics, Toronto, Canada (in-person and online), edition.

[2] Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, et al.. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

[3] Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating Multilingual Speech Translation under Realistic Conditions with Resegmentation and Terminology. In Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.

[4] Raymond Li, Wen Xiao, Lanjun Wang, Hyeju Jang, and Giuseppe Carenini. 2021. T3-Vis: visual analytic for Training and fine-Tuning Transformers in NLP. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 220–230, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.