Yuran Liu
Yongzhe Zhu
Yuxuan Zhu
Dec 2, 2021

CIS 530 Term Project Milestone 2

Direct Chinese to English Speech Translation with Sequence to Sequence Model

Evaluation

We feed our output speech (audio file) into an off-the-shelf English ASR system to produce a text form of our output. We use the same ASR system to produce a text form of the reference. Then, we compute the BLEU and NIST scores of the translation based on the above ASR transcriptions.

We use mteval-v13a, a script from the Moses project, to compute the individual N-gram (1-gram to 9-gram) and cumulate N-gram (1-gram to 9-gram) BLEU and NIST scores.We provided python scripts to convert audio files and json files into a sgm file which can be used for the evaluation script. We use a single reference for evaluation, and computes the brevity penalty based on the length of this reference. The reference and translation (in English) are tokenized by: spaces, punctuation(unless followed by digit), dash when preceded by digit.

Simple Baseline

We implemented a cascade system of ASR-MT-TTS using off-the-shelf APIs. We used Google speech recognition API in the SpeechRecognition python module for ASR. We used Google Translate API in the googletrans python module for MT. We used the Google TTS API in the gtts Python module. We cascaded the systems together to perform Speech-Speech translation.

The simple baseline was able to successfully generate an output wave file. From human perspective, the output sounded nice (synthesized wave); but the semantic meaning is not fully translated, and there were hilarious mistakes e.g. translating “大本营” (base camp) to “Big Ben” - “大本钟” in Chinese.

Strong Baseline

We implemented a text translation pipeline using pre-trained model by University of Helsinki on Chinese-English translation from HuggingFace (https://huggingface.co/Helsinki-NLP/opus-mt-zh-en). It is a transformer model with corpus from Opus, with SentencePiece pre-processing. The recorded BLEU score on its own testset, Tatoeba-test.zho.eng, is 36.1.

The pipeline worked well on pure translation, which is the core part of the proposed goal. The results were transferred to our evaluation script for deciding the performance, via record in JSON. The challenge on this model, though, is reflected in the inference time of the model. Translating a 200-sentence script paragraph cost

around 6 minutes, and the entire validation set would take ~90 min. Yet, it serves well as a baseline (we only need to infer once and compare). On the first large passage it analyzed, the BLEU score was ~18. Casual oral expression not being fully translated (filling words, exclamations, etc.,) might account for the slightly lower performance.