Correcting automatic speech recognition errors using pre-trained language models

Bc. Jiří Vlasák
xvlasa15@fit.vut.cz

Santosh Kesiraju Ph.D.
kesiraju@fit.vut.cz

Abstract

This project explores the use of language models to correct errors made by automatic speech recognition (ASR) systems. The chosen ASR system is the small variant of OpenAI Whisper and it is evaluated on the czech subset of Common Voice and Czech Parliament Plenary Hearings dataset. The results show that by employing a secondary language model to correct the outputs of a finetuned Whisper ASR system, the word error rate (WER) can be reduced from 23.6 to 21.4 for the Common Voice dataset and from 21.0 to 17.1 on both datasets combined. The improvement is even more pronounced when using ASR model that was not finetuned for the specific language or dataset. In this setting, the WER was reduced from 50.1 to 25.1 on the combined dataset when using the mBART-large-50 model to correct the errors. Despite these impressive results, qualitative analysis shows the limitations of this simple approach.

1 Task Definition

The task involves employing a secondary language model to enhance the accuracy of the ASR system's transcriptions for Czech speech. This correction process aims to reduce the Word Error Rate (WER) generated by the ASR system.

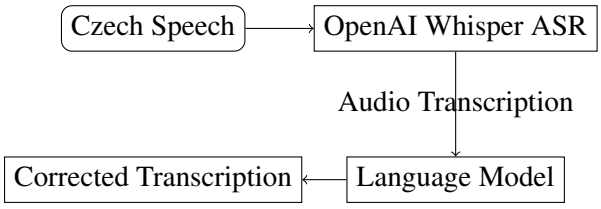

Figure 1: Diagram illustrating the process

2 Methodology

To obtain predictions first, we need an ASR dataset for Czech language. The Common Voice 13.0 (Ardila et al., 2019) and Czech Parliament Plenary Hearings (Kratochvil et al., 2020) datasets were chosen mainly for their respectable sizes and our experience with them from previous work.

OpenAI Whisper (Radford et al., 2023) was chosen as the ASR system to correct the predictions for. This was mainly done because we have experience with it from previous work finetuning it on Czech language. In this work we explore using both standard multilingual Whisper checkpoint, trained by OpenAI and our checkpoint which was finetuned on the dataset to provide better prediction for Czech language.

To provide corrections for the predictions done by the Whisper model, sequence to sequence natural language models are used. This work uses mT5 (Xue et al., 2020) and mBART (Liu et al., 2020) models. These models are trained on pairs of predictions from the ASR system and the original sentence.

To evaluate the method, the word error rate (WER) is computed for the predictions of the ASR system, as well as the predictions of the language models.

2.1 Data Collection

The Czech subset of the Common Voice dataset and the Czech Parliament Plenary Hearings dataset were utilized for this study. The data was preprocessed to extract audio samples and their corresponding transcripts. Common Voice dataset consist of both informal and formal language, while the parliament dataset consist of plenary hearings in usually more formal language.

Data from both datasets were combined to form the train and test set. The combined dataset has over 219 000 training utterances and 11 000 test utterances.

2.2 ASR System

The chosen ASR system is the small variant of OpenAI Whisper. This variant had the best results when

training on the combined dataset. We used both our finetuned checkpoint and the original checkpoint from OpenAI. The finetuned checkpoint was already trained on this dataset and had better performance on Czech language, while the original checkpoint was used to introduce more errors into the training data.

2.3 Language Model Integration

A language model was employed to correct errors produced by the ASR system. This model was utilized to enhance the accuracy of the ASR output by refining the transcriptions. This work utilized mT5-small and large variants as well as the mBART-large-50 model. The mT5 models are multilingual versions of the original T5 models. These models achieved SOTA results on many language tasks by converting the task to sequence-to-sequence form. The mBART models are multilingual versions of the BART models. They are primarly used for translation.

2.4 Evaluation Metrics

The primary evaluation metric used to assess performance enhancement was the Word Error Rate (WER). WER mainly measures how the model constructs syntactically and semantically sound sentences. The reduction in WER was measured to quantify the improvement achieved by integrating the language model into the ASR system.

3 Experimental Setup

The training was done on Cesnet Metacentrum cloud using various GPU servers. All of the models were implemented in Huggingface. The models were trained for 5 epochs with checkpoints for the best performing model on WER. The AdamW optimizer was used with learning rate of 1e-4 and beta parameters of 0.9 and 0.999. The language models usually achieved best performance after 1 or 2 epochs. After that their WER plateaued. All of the models were evaluated with using greedy search decoding.

4 Results and Analysis

At first our finetuned Whisper model `vtlustos/whisper-small` was ran on the Czech subset of the Common Voice 13.0 dataset to obtain the predictions for the language models to learn from. On pairs of these predictions and original transcriptions were trained the mT5-small, mT5-large and mBART-large-50 models.

Model	WER
vtlustos/whisper-small (finetuned ASR)	23.6
mT5-small	26.5
mT5-large	22.4
mBART-large-50	**21.4**

Table 1: Results on the czech subset of the Common Voice dataset when using already finetuned ASR model.

The results in table 1 show that the small mT5 model is not able to correct the mistakes and instead introduces even more errors and thus increasing the word error rate. The large mT5 model slightly improves to 22.4 over the baseline. This is probably mainly because of the larger learning capacity of the model. The mBART-large-50 model achieves the best results, decreasing the WER from 23.6 to 21.4. This is probably again the result of larger model capacity over the other two models.

Next, to introduce training data with more errors, the standard `openai/whisper-small` model was trained on the combined dataset of Common Voice and Czech Parliament Plenary hearings. This introduced training data with more errors, as the WER of the ASR system decreased to 50.1. The large mT5 model was not used this time.

Model	WER
openai/whisper-small (baseline ASR)	50.1
mT5-small	49.2
mBART-large-50	**25.1**

Table 2: Results on the combined dataset when using standard ASR model.

The results in table 2 show that the small mT5 model was again not able to substantially decrease the amount of errors in the transcribed text. However, the mBART-large-50 model significantly reduced the number of errors, dropping the WER from 50.1 to 25.1.

To get the full potential of this approach, the mBART-large-50 was then run on the predictions of the trained `vtlustos/whisper-small` model on the combined dataset. The finetuned model achieved WER of 21.0 compared to 50.1 of the standard model.

Sentence	Prediction	Correction
Až jindy.	Aší jindy.	A ší jindy.
Není nikterak rozjuchaný...	Není nikterak rozděuchaný...	Není nikterak rozzuchaný...
A končí to až ve dvě v noci.	A končí to až ve dvě v noci.	A končí to až ve dvě v noci.
Pane předsedající,...	Pane předsedající...	Pane předsedající,...
...zpráva, o níž zde hovoříme...	...zpráva, o niž zde hlořítе...	...zpráva, o níž zde hovoříte...
Jen mrknout!	Jenem rakela.	Je německý.

Table 3: Examples of the original transcriptions (Sentence), predictions made by the trained ASR system (Prediction) and the correction done by the mBART-large-50 model (Correction).

Model	WER
vtlustos/whisper-small (finetuned ASR)	21.0
mBART-large-50	**17.1**

Table 4: Results on the combined dataset when using finetuned ASR model and mBART-large-50 model trained on the predictions of baseline ASR model.

Table 4 shows the results of this approach. The mBART-large-50 model significantly decreases the WER of the finetuned ASR model, despite being trained on predictions from different checkpoint of the model. This can be, because the untrained ASR introduces more errors but similar type of errors as the trained ASR model.

4.1 Qualitative analysis

Qualitative analysis in table 3 show that despite significant decrease in WER, the language model struggles to correct mainly ambiguous predictions, where the original word cannot be easily deduced from the prediction. The Whisper ASR model frequently outputs a similar sounding sequence as the original, but often with small regard to czech language words and rules.

5 Conclusion

The utilization of language models to rectify errors generated by automatic speech recognition (ASR) systems has shown promising results in enhancing transcription accuracy for Czech speech. This project explored the use of a secondary language model, specifically the mT5 and mBART models, to refine and correct the outputs of the ASR system, significantly reducing the Word Error Rate (WER).

The findings demonstrated notable improvements in WER, showcasing a decrease from 23.6 to 21.4 on the Czech subset of the Common Voice dataset and from 21.0 to 17.1 on a combined dataset of Common Voice and Czech Parliament Plenary Hearings.

Despite these impressive outcomes, qualitative analysis highlighted certain limitations. The language model struggled notably when correcting ambiguous predictions or instances where the original word couldn't be easily inferred from the ASR output.

While the integration of a secondary language model proved highly beneficial in reducing WER for Czech ASR, challenges persist in addressing ambiguous and linguistically complex speech segments. Training the language model on more Czech data could help to make the corrector model more robust. Including the audio embeddings as additional input could also be beneficial.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Jonáš Kratochvil, Peter Polák, and Ondřej Bojar. 2020. Large corpus of Czech parliament plenary hearings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6363–6367, Marseille, France. European Language Resources Association.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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*, pages 28492–28518. PMLR.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.