Fine-Tuning Pre-Trained Language Model for Crowdsourced Texts Aggregation

Mikhail Orzhenovskii[1]

Abstract

We report on our system for aggregating crowdsourced texts for the VLDB 2021 Crowd Science Workshop's shared task. In the task, for each original audio, several crowdsourced transcriptions need to be combined into a single transcription. We propose a system that uses a pre-trained language model, fine-tuned on the augmented dataset, and task-specific post-processing of the model's outputs to improve the quality of the results. Our model scored 95.73 (45% fewer mistakes compared to the baseline) and achieved the 1^{st} place on the shared task leaderboard. [1]

Keywords

Crowdsourcing, Text aggregation, Truth discovery

1. Introduction

VLDB 2021 Crowd Science Challenge[1] is a shared task on aggregation of crowdsourced texts. Multiple transcriptions made by people needed to be aggregated to produce a single high-quality transcription. The audios were produced using a voice assistant from Wikipedia articles.

The problem is that some annotators can be unskilled or malicious. One more thing, different people can make mistakes in different parts of the sentence. The data is very noisy.

The metric used to evaluate the solutions in the shared task was highest Average Word Accuracy (AWAcc). Word Accuracy is calculated as

$$WAcc = 100 \times max(1 - WER, 0)$$

This aggregation task can be seen as a particular case of multi-document summarization or as mistake correction. Pre-trained language models are widely used for many text-related tasks, including text summarization. Linguistic knowledge is beneficial in this task because it helps to choose the possible word sequences, or replace a misheard word with a word with high probability in the context. We applied end-to-end training because the available dataset was large enough.

[1]Source code is available on https://github.com/orzhan/bart-transcription-aggregation

VLDB 2021 Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale, August 20, 2021, Copenhagen, Denmark

Table 1
Example of data (id 8359)

Transcriptions:	`Her recent|Her research interests include; number theory, Houdin theory automorphic forms and spectral graph theory.|Her research interests include number Theory coding Theory Autumn orphic forms and spectral graph Theory |research interests include number theory coding theory automorphic forms and spectral graph theory|her research interests include number theory coding theory automorphic forms and spectural graph theory|Her research interests include ,number theory, padding theory,automorfic forms and spectual graph theory|,mnlknk`
Ground truth:	`her research interests include number theory coding theory automorphic forms and spectral graph theory`

2. Related work

ROVER system used dynamic programming to align and augment word transition networks. After joining the networks, the final WTN was searched by the scoring module to select the best sequence[2].

In HRRASA system, multiple crowdsourced sequences were aggregated using global annotator reliability and local question-wise reliability based on text similarities[3].

3. Dataset

For each of 9700 task ids, the training dataset contained 7 transcriptions made by the annotators and the ground truth text. The testing dataset contained 4502 task ids with 7 transcriptions for each id.

The ground truth texts were typically from 8 to 15 words long. The number of different words used in transcriptions was 1 to 4 times larger than the number of different words in the ground truth label. This indicates that some texts were easier for the annotators than the other ones. An example of the data is shown in Table 1.

4. Model and post-processing

Text aggregation can be seen as a sequence-to-sequence task: the input sequence is a concatenation of the crowdsourced transcriptions separated with a delimiter. The output sequence is the ground truth text. The order of transcriptions does not matter, and all of them can be treated equally, so we generated four sequences with different orders of transcriptions for each task id. This method partially helped to regularize the model.

We have evaluated two pre-trained language models: T5[4] and BART[5]. Both models use the same encoder-decoder architecture and are capable of solving sequence-to-sequence problems.

The evaluation metric in the shared task was based on Word Error Rate, making, for example, *color* and *colour* different words. In the training dataset's ground truth labels, American English forms were more frequent, so we converted the model's outputs from British English to American

English (if applicable) with vocabulary from American British English Translator[1].

Shuffling the transcriptions helped the model to regularize; however, it was sometimes sensitive to the order of the inputs. To obtain more stable results for the test dataset, for each task id, we inputted 20 concatenations with different orders of transcriptions and selected the final result using a majority vote. For most examples, there were only two different generated results, one of which outputted for most of the 20 concatenations. The input permutations were chosen to maximize the total Kendall tau rank distance between them.

5. Experiments

For the experiments we were using transfomers[6] and simpletransformers[7] libraries which support both BART and T5 models. The models were pre-trained on different tasks (summarization and translation), so fine-tuning was necessary to use them in the aggregation task. We fine-tuned the pre-trained models on 9400 samples of the training dataset. Another 300 samples were used as evaluation dataset to choose the training parameters and to select the best model.

T5 model was producing nearly the same results as BART model, but the fine-tuning was taking about 4 times longer, so we chose BART and did the most experiments with it. As expected, the larger models outperformed the smaller ones, so BART-large was selected for the final experiments.

We selected a relatively small base learning rate 4×10^{-6}, and followed transformers' default schema of changing learning rate during fine-tuning (Fig. 5). Batch size during training and evaluation was set to 8 as it was the maximum size fitting on the GPU.

We stopped training after the 5th epoch when evaluation AWAcc stopped to increase (Fig. 5). Evaluation loss started to rise during the 1st epoch, but further training helped obtain higher WER scores on evaluation and public test datasets. The increase of the evaluation loss can indicate over-fitting, but the actual target metric WER is different and is not always correlated with the evaluation loss (which is based on maximum likelihood, not error count).

Beam search with 5 beams slightly improved the score compared to greedy decoding. Using more beams did not lead to better results.

6. Results

The results of the model on the different datasets are shown in the Table 2. The difference between the evaluation score and public/private scores is relatively small.

The results of the proposed model and the baselines are shown in the Table 3. Majority vote stands for selecting the most common result from the transcriptions. Random choice stands for choosing a random transcription as the answer.

Examples of the model's outputs are displayed in the Table 4. The model processed 73.14% of the inputs without any error, the first two examples belong to this group. The other 26.86% of the inputs contained some mistakes, which is illustrated by the third example.

[1] https://github.com/hyperreality/American-British-English-Translator

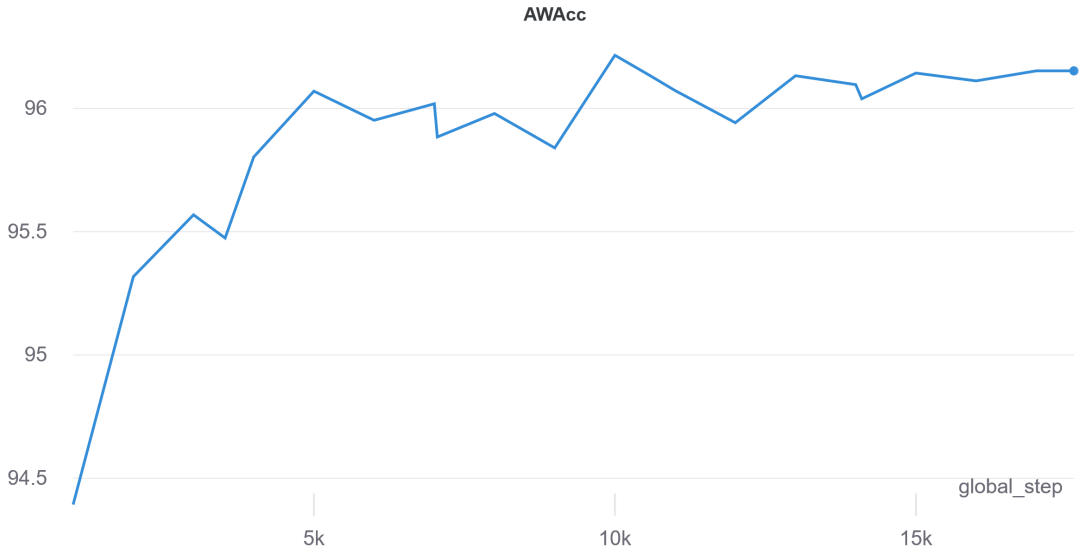

Figure 1: Evaluation AWAcc

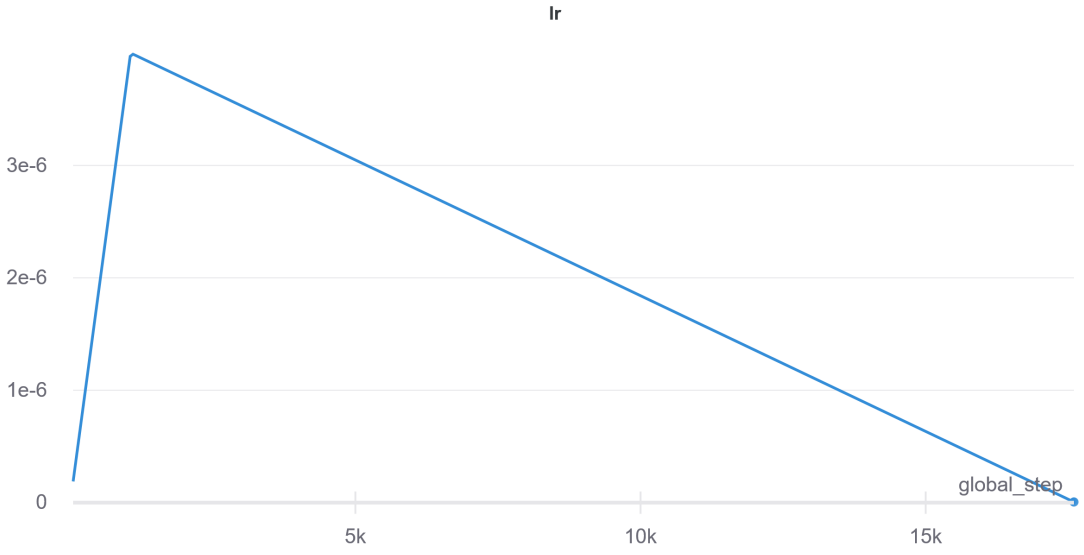

Figure 2: Learning rate

7. Conclusion

The proposed model outperformed the benchmark and other models, achieving a high score of 95.73 on the shared task. The model only used the texts of the transcriptions (no information about the annotators) to produce the result.

Figure 3: Training loss

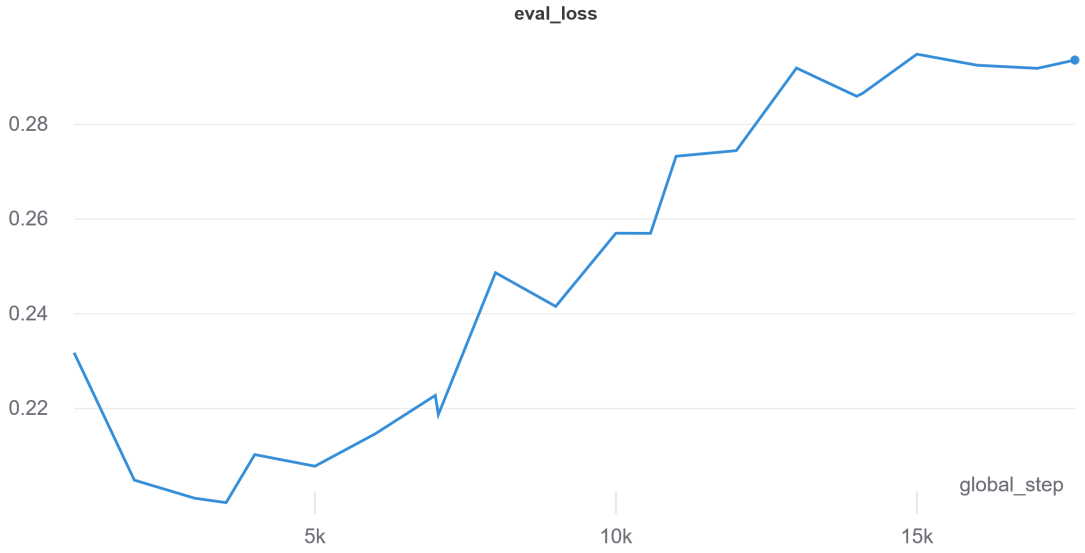

Figure 4: Evaluation loss

Possible improvements in quality can be achieved by using the information about the annotators (for example to assign higher weights to accurate annotators), injecting phonetic knowledge into the model to match misheard word sequences better, or using symmetric model architecture that processes input transcriptions in parallel (removing the need of permutations during training and inference).

Table 2
AWAcc of the final model

Dataset	Score
evaluation set	96.10
public test	95.75
private test	95.73

Table 3
AWAcc on private test compared to baselines

Model	AWAcc
Final model	95.73
Final model (no shuffling)	95.55
ROVER baseline[2]	92.25
HRRASA baseline[3]	91.04
Majority vote	72.42
Random choice	68.75

References

[1] D. Ustalov, N. Pavlichenko, I. Stelmakh, D. Kuznetsov, VLDB 2021 Crowd Science Challenge on Aggregating Crowdsourced Audio Transcriptions, in: Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale, Copenhagen, Denmark, 2021.

[2] J. Fiscus, A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover), in: 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, 1997, pp. 347–354. doi:10.1109/ASRU.1997.659110.

[3] J. Li, Crowdsourced text sequence aggregation based on hybrid reliability and representation, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1761–1764. URL: https://doi.org/10.1145/3397271.3401239. doi:10.1145/3397271.3401239.

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[5] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. arXiv:1910.13461.

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics,

Table 4
Examples of results

Example of an easy aggregation **Transcriptions:** the jungle finally offering some protection|the jungle finally offering some protection|the jungle finally offering some protection|the jungle finally offering some protection|the jungle finally offering some protection |the jungle finally offering some protection |the gentil finally offering some protection **Ground truth:** the jungle finally offering some protection **Prediction:** the jungle finally offering some protection **AWAcc:** 100.0
Example of a difficult aggregation done correctly **Transcriptions:** her voice caught only underwear |her voice confirmed the word|her voice caught only once|a voice coach on the word|her voice got only word|her voice called on the word|he voice caused only worries **Ground truth:** her voice caught on the word **Prediction:** her voice caught on the word **AWAcc:** 100.0
Example of an incorrect aggregation **Transcriptions:** an anger rich in pain n dissolution |anf im underated and ojo|un anger rated in pain and disillusionment|and i deleted in pain and desolutionment|an angry richard in pain and disillusionment|thanks for the heads are|an anger routed in pain and disillusionment **Ground truth:** an anger rooted in pain and disillusionment **Prediction:** an anger rich in pain and disillusionment **AWAcc:** 85.71

Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[7] T. C. Rajapakse, Simple transformers, https://github.com/ThilinaRajapakse/simpletransformers, 2019.