VLDB 2021 Crowd Science Challenge on Aggregating Crowdsourced Audio Transcriptions

Dmitry Ustalov[1], Nikita Pavlichenko[2], Ivan Stelmakh[3] and Dmitriy Kuznetsov[2]

[1]Yandex, Piskariovski Prospekt, building 2, block 2, Benois Business Centre, Saint Petersburg, 195027, Russia
[2]Yandex, Ulitsa Lva Tolstogo 16, Moscow, 119021, Russia
[3]Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213, USA

Abstract

The growing interest in automatic speech recognition methods urges new high-quality datasets and benchmarks for reproducible and reliable evaluation. Crowdsourcing has become an efficient way for audio transcriptions in real-world applications, such as call centers and voice assistants. However, as the recordings' difficulty and the expertise of the crowd workers vary, each recording is usually transcribed by multiple workers, raising the need for principled aggregation methods for word sequences from multiple noisy inputs. This paper reviews the crowdsourced audio transcription shared task devoted to this problem and co-organized with the Crowd Science Workshop at VLDB 2021; the competition attracted 18 participants, 8 of them have successfully beaten our non-trivial baselines. We describe the task dataset, evaluation criterion, competition timeline, as well as the proposed baselines and participating systems. The winning systems treated the difficult crowdsourced sequence aggregation task as the better-studied text summarization task, enabling fine-tuned large-scale language models to establish the new state-of-the-art in this problem.

Keywords

audio transcription, crowdsourcing, aggregation, speech generation

1. Introduction

The problem of automatic speech recognition arises in various domains, spanning from voice assistants to call centers and accessibility tools. Although the state-of-the-art machine learning models like the Conformer or its derivatives [1] show impressive progress in recognizing spoken language, we believe in the urgency of creating additional, more challenging datasets aimed at less popular natural languages, speaker backgrounds, and domains. Even though crowdsourcing has become a popular approach for collecting audio transcriptions to evaluate or augment the speech recognition methods [2], the high variance of the recordings quality and crowd worker skills urges better quality control techniques for crowdsourced audio transcriptions. To account

VLDB 2021 Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale, August 20, 2021, Copenhagen, Denmark
dustalov@yandex-team.ru (D. Ustalov); pavlichenko@yandex-team.ru (N. Pavlichenko); stiv@cs.cmu.edu (I. Stelmakh); wth-dmitriy@yandex-team.ru (D. Kuznetsov)
https://linkedin.com/in/ustalov/ (D. Ustalov); https://linkedin.com/in/nikita-pavlichenko/ (N. Pavlichenko); https://www.cs.cmu.edu/~istelmak/ (I. Stelmakh); https://linkedin.com/in/dmitriy-kuznetsov-4997b8167/ (D. Kuznetsov)
0000-0002-9979-2188 (D. Ustalov); 0000-0002-7330-393X (N. Pavlichenko)

Table 1
Shared Task Dataset Overview

Dataset	# of Recordings	# of Transcriptions	# of Workers
Train	9,700	67,900	1,879
Test	4,502	31,514	1,160
Test (Public Only)	1,500	10,500	1,039
Test (Private Only)	3,002	21,014	1,127

for this problem, each audio recording is typically transcribed into a sequence of words by multiple crowd workers. But *how do we aggregate these multiple transcriptions* to obtain the final high-quality transcription without using the post-acceptance mechanism [3]?

Answer aggregation is well-studied for categorical responses [4], but one cannot trivially apply such methods to sequences. We ran an open competition, aka shared task, in conjunction with the VLDB 2021 Crowd Science Workshop to answer this question.[1] Specifically, the goal of the participants is to build a model that aggregates multiple transcriptions of audio obtained on a crowdsourcing platform into a single high-quality transcription.

This paper reports the results of the shared task on aggregating crowdsourced audio transcriptions that attracted 18 participants, 8 of them have successfully beaten our non-trivial ROVER baseline. The paper is organized as follows. Section 2 describes the dataset offered at the shared task and the average word accuracy evaluation criterion. Section 3 presents the competition setup and timeline. Section 4 overviews the baselines, the winning systems, and some of the prominent participating systems. Section 5 concludes with the final remarks.

2. Dataset and Evaluation Criterion

We used the Vox DIY approach to produce an English language dataset for our shared task [5]. We sampled and normalized sentences from English Wikipedia and BookCorpus, and synthesized audio recordings using the Yandex SpeechKit text-to-speech service.[2] After that, we ran annotation of these recordings on the Toloka crowdsourcing platform in which every recording received at least five transcriptions from different workers.[3] In total, we obtained 99,414 transcriptions for 9,700 recordings submitted by 3,039 workers. Finally, we split the recordings and the received transcriptions into the train, public test, and private test datasets comprising the whole shared task dataset. Table 1 shows a detailed description of our dataset.

Since our dataset includes the ground truth transcription for each recording, we decided to apply the traditional speech recognition quality criterion called *word error rate* (WER). Given the number of substitutions (S), deletions (D), insertions (I), correct words (C), the word error rate is defined as follows [6]:

$$\mathrm{WER} = \frac{S + D + I}{S + D + C}$$

[1] https://crowdscience.ai/challenges/vldb21
[2] https://cloud.yandex.com/en/services/speechkit
[3] https://toloka.ai/

We used the well-known JiWER library for Python to compute the word error rate per recording for compatibility and reproducibility.[4] Then, we aggregated these scores into *average word accuracy* (AWAcc), which we used as the evaluation criterion for our competition:

$$\mathrm{AWAcc} = \frac{1}{|R|} \sum_{r \in R} \max\left(0, 1 - \mathrm{WER}(r)\right) \times 100\%$$

3. Competition

We hosted our competition on the Yandex.Contest platform. There were two distinct phases: Practice and Evaluation.[5] At the former phase, the participants evaluated their methods on the same publicly available train dataset with the known ground truth. We made this phase to allow the participants to get used to the competition format, data, and platform. At the latter phase that has been opened later, the participants had to submit their predictions on the test dataset without knowing the ground truth. The scores were computed only on the public test dataset during the competition, while the final standings were computed on the private part. The three systems with the highest AWAcc scores are considered the winners. After the competition, we invited the participants to submit their system description papers to the VLDB 2021 Crowd Science Workshop.[6] These papers undergo peer review along with the regular papers submitted to the workshop. As a result, the competition had the following timeline:

- Practice Phase Started: April 15, 2021
- Evaluation Phase Started: May 5, 2021
- Evaluation Phase Finished: June 18, 2021
- System Description Paper Deadline: July 5, 2021
- VLDB 2021 Crowd Science Workshop; August 20, 2021

4. Systems

In this section, we describe the winning systems and offer a few details about some of the prominent participating systems. Our competition has attracted 18 different participants. The complete final standings are shown in Appendix A.

4.1. Baselines

We offered two kinds non-trivial baselines in our shared task, ROVER and RASA & HRRASA.

ROVER. The first baseline is Recognizer Output Voting Error Reduction (ROVER) that uses dynamic programming to align the input sequences and outputs a new word sequence using a majority vote on each token [7]. Although it was initially designed for aggregating the results of multiple automatic speech recognition methods, it can also successfully handle crowdsourced

[4] https://github.com/jitsi/jiwer/

[5] https://contest.yandex.com/contest/27051/enter and https://contest.yandex.com/contest/27274/enter

[6] https://crowdscience.ai/challenges/vldb21

transcriptions [2]. An important property of ROVER is its ability to produce a transcription that differs from every transcription submitted by the workers. On the private part of our test dataset, ROVER showed the average word accuracy of 92.25%.

RASA & HRRASA. The second baseline is Reliability Aware Sequence Aggregation (RASA) and its modification called HRRASA [8]. Both methods use a large-scale language model to select the best transcription based on mean weighted embedding iteratively. HRRASA additionally takes into account the local reliabilities represented by the GLEU distance between the particular response to other responses for the recording. Unlike ROVER, the output of either RASA & HRRASA is always one of the submitted transcriptions. On the private part of the test dataset, both RASA and HRRASA showed the average word accuracy of 91.04%.

4.2. Winning Systems

Out of the 18 participants of our competition, only 8 of them have beaten our ROVER baseline. The winning system has achieved the impressive AWAcc of 95.73% on the private subset of our data, i.e., their model makes almost twice as few mistakes comparing to the baseline that showed the AWAcc of 92.25%. Among the three winning systems, only one has taken into account the worker features.

1st Place: Fine-Tuned Text Summarization. This participant used a pre-trained language model for text summarization, fine-tuned on the augmented shared task dataset [9]. During the hyper-parameter search, the best results were shown by the BART model [10]. The augmentation of shuffling the input transcriptions allowed to regularize the model. As a post-processing step, they replaced British English word forms with those from American English, e.g., colour → color. This system showed the AWAcc of 95.73% on the private test dataset.

2nd Place: Text Summarization. This participant used a similar approach to the 1st place, yet they did not perform data augmentation [11]. Having tried different Seq2Seq models, they found that the T5 model demonstrates the best results [12], while using additional external datasets does not improve the results. This system showed the AWAcc of 95.66% on the private test dataset.

3rd Place: Model Combination and Toloka. This participant built a linear combination of mean WER between the current hypothesis and others, pre-trained BERT language model bert-base [13], hypothesis classifier tuned from the pre-trained language model for one epoch, hypothesis length, and two worker features: annotation consistency and the total number of responses. Additionally, they have improved the results by collecting 1.5K phrases from open sources using Toloka by asking the workers to find the best hypothesis for the model on the first page of the Yandex search results.[7] This system showed the AWAcc of 95.48% on the private test dataset.

[7] https://yandex.com/

4.3. Honorable Mentions

We believe that certain approaches are also worth mentioning in our shared task organization report.

4th Place: Fine-Tuned Language Model and Gradient Boosting. This participant used the T5 model fine-tuned on the shared task dataset; for augmentation, they have shuffled the order of candidates. Then, they fitted a LightGBM model [14] to rank the T5 output and the original set of candidates, which improved the score. This system showed the AWAcc of 95.00% on the private test dataset.

6th Place: Levenshtein-Median. This participant used an elegant approach of choosing each word to be the closest to the median Levenshtein distance to other words. Adding the worker weights and excluding those who tend to submit inconsistent results allowed improving the score. This system showed the AWAcc of 93.37% on the private test dataset.

5. Conclusion

For us, the key takeaway from the organized shared task was the possibility to treat the crowdsourced text aggregation task as a well-known text summarization task. Using large-scale language models allowed one to improve over the ROVER dynamic programming baseline method proposed almost thirty years ago. The winning system reduced the average WER almost twice from 7.75% of this baseline to 4.27%. Another notable insight was the opportunity to compute a simple Levenshtein median also shows meaningful results, but this approach requires a careful selection of workers. Having released everything from the raw texts to the baseline methods and scoring program under a permissive license,[8] we believe that the outcomes of this competition will help in the creation of general-purpose quality control techniques in open-ended crowdsourcing tasks besides audio transcriptions.

Acknowledgments

We would like to thank all the shared task participants for advancing the state-of-the-art in this challenging audio transcription problem. We are grateful to the Toloka team for providing the crowdsourcing budget. We are also thankful to the Yandex.Contest team for helping us during the competition.

References

[1] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, Conformer: Convolution-augmented Transformer for Speech Recognition, in: Proc. Interspeech 2020, International Speech Communication Association, Shanghai, China, 2020, pp. 5036–5040. doi:10.21437/Interspeech.2020-3015.

[8] https://github.com/Toloka/VLDB2021_Crowd_Science_Challenge

[2] M. Marge, S. Banerjee, A. I. Rudnicky, Using the Amazon Mechanical Turk for transcription of spoken language, in: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2010, IEEE, Dallas, TX, USA, 2010, pp. 5270–5273. doi:10.1109/ICASSP.2010.5494979.

[3] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, K. Panovich, Soylent: A Word Processor with a Crowd Inside, in: Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology, UIST '10, ACM, New York, NY, USA, 2010, pp. 313–322. doi:10.1145/1866029.1866078.

[4] A. Sheshadri, M. Lease, SQUARE: A Benchmark for Research on Computing Crowd Consensus, in: First AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2013, Association for the Advancement of Artificial Intelligence, 2013, pp. 156–164. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/13088.

[5] N. Pavlichenko, I. Stelmakh, D. Ustalov, CrowdSpeech and Vox DIY: Benchmark Dataset for Crowdsourced Audio Transcription, in: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 2021. URL: https://openreview.net/forum?id=3_hgF1NAXU7. arXiv:2107.01091.

[6] L. R. Bahl, F. Jelinek, Decoding for Channels with Insertions, Deletions, and Substitutions with Applications to Speech Recognition, IEEE Transactions on Information Theory 21 (1975) 404–411. doi:10.1109/TIT.1975.1055419.

[7] J. G. 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, IEEE, Santa Barbara, CA, USA, 1997, pp. 347–354. doi:10.1109/ASRU.1997.659110.

[8] 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, Virtual Event, China, 2020, pp. 1761–1764. doi:10.1145/3397271.3401239.

[9] M. Orzhenovskii, Fine-Tuning Pre-Trained Language Model for Crowdsourced Texts Aggregation, in: Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale, Copenhagen, Denmark, 2021.

[10] 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, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. doi:10.18653/v1/2020.acl-main.703.

[11] S. Pletenev, Noisy Text Sequences Aggregation as a Summarization Subtask, in: Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale, Copenhagen, Denmark, 2021.

[12] 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: https://jmlr.org/papers/v21/20-074.html.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), NAACL-HLT 2019, Association for Computational Linguistics, Minneapolis, MN, USA, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree, in: Advances in Neural Information Processing Systems 30, NIPS 2017, Curran Associates, Inc., 2017, pp. 3149–3157. URL: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

A. Competition Results

Table 2 shows the competition results.

Table 2
Final Standings in the Competition

Place	AWAcc, %	
	Public	Private
1	95.75	95.73
2	95.67	95.66
3	95.62	95.48
4	95.20	95.00
5	94.55	94.14
6	93.07	93.37
7	92.60	92.54
8	92.19	92.47
ROVER		92.25
9	91.65	91.09
(HR)RASA		91.04
10	90.12	90.42
11	89.67	90.35
12	90.37	90.29
13	79.51	79.46
14	78.72	78.70
15	78.72	78.34
15	78.72	78.34
16	75.59	76.01
17	1.37	0.00