The 6th International Conference on Interdisciplinary research on Computer science, Psychology, and Education (ICICPE' 2022)
December 26-28, 2022. Pattaya, Thailand.

Automatic Generation of Learning Data for Korean Speech Recognition Post-Processor

Seonmin Koo[1], Chanjun Park[1,2], Hyeonseok Moon[1], Jaehyung Seo[1],
Sugyeong Eo[1], Heuiseok Lim[1]

[1]Korea University, Computer Science, [2]Upstage
{fhdahd, bcj1210, glee88, seojae777, djtnrud, limhseok}@korea.ac.kr

Abstract. To train the post-processor, we need a parallel corpus containing the error types. One simple way to do this is to insert an error into the correct answer sentence and generate an error sentence to create a pseudo-parallel corpus. However, there is a possibility that this is not an actual error. To alleviate this, there is a methodology for generating a parallel corpus for post-processor model training using Back TranScription (BTS). However, there is a point of view that noise may be less when generated by the method. Therefore, in this study, the performance of the BTS method and the method with artificially added noise intensity are compared.

Keywords: Deep Learning, Natural Language Processing, Speech Recognition Post-Processor, Error Insertion, Parallel Corpus Creation

1 Introduction

Recently, as research in automatic speech recognition research and its application to actual commercialization systems has become active, research to improve the accuracy and performance of automatic speech recognition systems is steadily progressing. As one of the studies to increase the accuracy of speech recognition, a study on a speech recognition post-processor model has been conducted. In order to train a post-processor model, a parallel corpus containing high-quality error types is required. Low-resource language datasets such as Korean have fewer data sets than high-resource languages. In order to alleviate this problem, existing studies use Back TranScription (BTS) to build a high-quality parallel corpus without human intervention.

Back TranScription is converting single text data into voice data using TTS (Text-to-Speech) and STT (Speech-to-Text) technologies, which are speech synthesis technologies, and then converting the converted voice data back to text. It is a methodology that generates parallel data with actual noise through. The method may have a view that the performance of TTS and STT is outstanding so that a parallel corpus containing only low-intensity noise can be generated. Therefore, in this paper,

a parallel corpus with increased noise intensity by artificially inserting errors is generated and used as a comparison group for the BTS methodology. The performance is compared by training each model with the parallel corpus generated by each methodology. Qualitative analysis results showed that BTS helps the speech recognition post-processor performance by including more practical errors that may occur in actual speech recognition situations.

2 Generate Noise Parallel Data

The BTS generates a voice file from a single corpus containing almost no noise through the TTS technology and then generates data through the STT technology for the file[2]. As it passes through TTS and SST, data containing errors in speech recognition results are generated for a single corpus. Spelling and grammar errors such as spacing errors and punctuation errors are mainly generated in the generated error types. Data generated using the method corresponds to a source sentence among the parallel corpus, and a single corpus used for generation corresponds to a correct sentence. By performing sequence-to-sequence model training using the data generated through the BTS, a high-quality speech recognition post-processor can be generated.

For comparison with BTS, we generate parallel data through a predefined error insertion methodology[3]. Specifically, a single corpus is used as an answer sentence, and an error type is inserted into the sentence to generate a raw sentence, resulting in a parallel corpus with a stronger noise intensity than BTS. Considering the real-world speech recognition environment, we generate data that performed error insertion, including several predefined ones, to generate a parallel dataset with stronger noise.

3 Qualitative Analysis

In this study, qualitative analysis was performed to practically analyze the reason for the performance difference depending on the methodology for generating the dataset for post-correction model training. Qualitative analysis was performed on the BTS model and the model with the best performance among the predefined error insertion models. Table 1 shows the actual results of error correction for each model.

Table 1. Examples of the results of correction creation for each model

Source sentence	하늘거리는 쉬폰 소재라 내 피부에 닿을 때마다 마음이 간지러워요 자 설레는 느낌이 들거든요 (It's made of a fluttering chiffon material, so whenever it touches my skin, my heart itches. It makes my heart)
Target sentence	하늘거리는 시폰 소재라 내 피부에 닿을 때마다 마음이 간지러워져 설레는 느낌이 들거든요. (It's a fluttering chiffon material, so whenever it touches my skin, my heart tickles and I feel excited.)
Noised Model	하늘거리는 시폰 소재라 네 피부에 닿을 때마다 마음이 간지러워요. 자 설레는 느낌이 들거든요. (It's made of fluttering chiffon material, so my heart tickles every time it touches my skin. I feel excited.)
BTS	하늘거리는 시폰 소재라 내 피부에 닿을 때마다 마음이 간지러워져서 설레는 느낌이 들거든요. (It's a fluttering chiffon material, so whenever it touches my skin, my heart tickles and I feel excited.)
Source sentence	그들은 태 군부대에 대한 올바른 생각을 가져야 하고 자신의 위치와 임무에 최선을 다해야 해요 (They need to have the right idea of the Tae army and do their best in their positions and missions)

Target sentence	그들은 타군부대에 대한 올바른 생각을 가져야 하고, 자신의 위치와 임무에 최선을 다해야 해요. (They should have the right idea of other military and do their best in their positions and missions.)
Noised Model	그들은 태국 공부에 대한 올바른 생각을 가져야 하고 자신의 위치와 임무에 최선을 다해야 해요. (They should have the right idea of studying Thailand and do their best in their positions and missions.)
BTS	그들은 타 군부대에 대한 올바른 생각을 가져야 하고 자신의 위치와 임무에 최선을 다해야 해요 (They need to have the right idea of other military units and do their best in their positions and missions)

In order to convey the meaning of a sentence well in an actual speech recognition situation, it is necessary to distinguish the end of a sentence. In this case, the BTS model was calibrated correctly, but the Noised Model was not calibrated with 'Gree Main'. From this, it can be seen that the BTS model distinguishes the endings of sentences better than the simple rule-based insertion method.

Even for words that are frequently used in the process of post-correction of recognition results, it is important to classify appropriate cases and correct them. For example, in the second sentence of Table 1, Noised Model incorrectly corrected 'Tae army' to 'Thailand', but BTS properly corrected it to 'other military'. This shows that the BTS methodology helps to perform post-correction well because it contains practical errors.

Due to the characteristics of automatic voice technology widely applied to actual commercial systems, the quality of voice recognition is one of the important factors to increase user satisfaction. Through the experimental results, we show that the BTS methodology helps to perform post-hoc correction well because it contains practical errors.

4 Conclusion

In this study, through comparative and qualitative analysis, the BTS methodology not only has the best quantitative performance, but also contains practical errors, so it is shown that it is an appropriate methodology to apply to an actual speech recognition system.

5 Acknowledgement

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00369, (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge).“This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)”

References

1. Paksoy, T., Kochan, C. G., & Ali, S. S. (Eds.). (2020). Logistics 4.0: digital transformation of supply chain management. CRC Press.
2. Park, C., Seo, J., Lee, S., Moon, H., Eo, S., & Lim, H. (2021). A Study on Verification of Back TranScription (BTS)-based Data Construction. Journal of the Korea Convergence Society, 12(11), 109-117.
3. Park, C., Kim, K., Yang, Y., Kang, M., & Lim, H. (2021). Neural spelling correction: translating incorrect sentences to correct sentences for multimedia. Multimedia Tools and Applications, 80(26), 34591-34608.