International Journal of Advanced Smart Convergence Vol.12 No.4 171-176 (2023)
http://dx.doi.org/10.7236/IJASC.2023.12.4.171

IJASC 23-4-21

A Study on DNN-based STT Error Correction

Jong-Eon Lee

Professional, Enterprise Division, LG UPLUS
jongeonlee@lguplus.co.kr;medosecy@gmail.com

Abstract

This study is about a speech recognition error correction system designed to detect and correct speech recognition errors before natural language processing to increase the success rate of intent analysis in natural language processing with optimal efficiency in various service domains. An encoder is constructed to embedded the correct speech token and one or more error speech tokens corresponding to the correct speech token so that they are all located in a dense vector space for each correct token with similar vector values. One or more utterance tokens within a preset Manhattan distance based on the correct utterance token in the dense vector space for each embedded correct utterance token are detected through an error detector, and the correct answer closest to the detected error utterance token is based on the Manhattan distance. Errors are corrected by extracting the utterance token as the correct answer.

Keywords: *STT Error Correction, Siamese Neural Networks, LSTM, DNN, CNN*

1. Introduction

Generally, voice-based services include a voice recognition process that receives the speaker's voice as input and outputs natural language text, and a natural language processing process that analyzes the output natural language text to extract the speaker's intention, emotion, and metadata. Afterwards, depending on the service, an answer is generated and provided as text or an audio answer is provided through voice synthesis. A speech recognition system consists of an acoustic model and a language model. An acoustic model analyzes the characteristics of speech signals, and a language model predicts words and phonemes through the recognized phonemes and the probability of occurrence between elements before and after the word. Acoustic models and language models require performance improvement through continuous learning and construction of similar pronunciation dictionaries within the speech recognition system to increase speech recognition accuracy. However, despite improved speech recognition performance, speech recognition errors continue to occur due to pronunciation variations and exceptional pronunciation laws.

Although a pronunciation dictionary is built to supplement voice recognition performance, there are limitations in continuously updating the dictionary for service terms, new words, and colloquialisms that are continuously created. In other words, voice recognition outputs selected text through its own language model

Manuscript Received: october. 18, 2023 / Revised: october. 24, 2023 / Accepted: october. 29, 2023
Corresponding Author: jongeonlee@lguplus.co.kr;medosecy@gmail.com
Tel: +82-10-8719-1099
Professional, Enterprise Division, LG UPLUS, Korea

and acoustic model, but the text selected and output also produces errors. These errors in sound and pronunciation-based voice recognition engine frequently occur in product name and service name domains where new words and compound nouns are frequent. In this case, the acoustic model and language model of the voice recognition engine must be retrained, or a dictionary of similar pronunciations must be used. Services are being provided by manually updating and supplementing the service, but this is inefficient in terms of manpower and resources.

Therefore, in this paper, we propose a speech recognition error correction system that can improve natural language processing results in various domains by detecting and correcting speech recognition errors in advance.

2. Related Works

The speech recognition error correction system proposed in this paper uses Siamese Neural Networks and bi-directional LSTM to ensure that one or more error speech tokens corresponding to the correct speech token are all located in a dense vector space for each correct speech token with similar vector values.

- **Siamese Neural Networks.** Siamese Neural Networks are a class of artificial neural network structures that have two or more identical sub-networks. Multiple sub-networks have the same parameters and weights and can be updated in synchronization with each other. This network has an effective structure for learning and mapping symmetric information matching results [1].
- **LSTM.** LSTM is a model that was created to solve the long-term dependencies of RNN. Therefore, it is a model that was created to predict future data by considering not only the previous data, but also past data on a more macroscopic level. LSTM learns by dividing short-term memory and long-term memory, and then merges the two memories to predict event probability [2].
- **Kanishka Rao's study.** In this study, Siamese Neural Networks and LSTM were used to learn input document pairs of variable length and create a distributed representation of documents that can more accurately render the semantic distance between document pairs. As a result, documents related to the same semantic or topic label were mapped to similar expressions with higher semantic similarity [3].

3. Design and Implementation

The processing procedure of the proposed system is shown in Figure 1. The uttered speech is converted into text and transmitted to NLU. If there is an error in the uttered speech and the intention analysis fails, it is sent back to the NLU after error detection and error correction, and the intention analysis is attempted again.

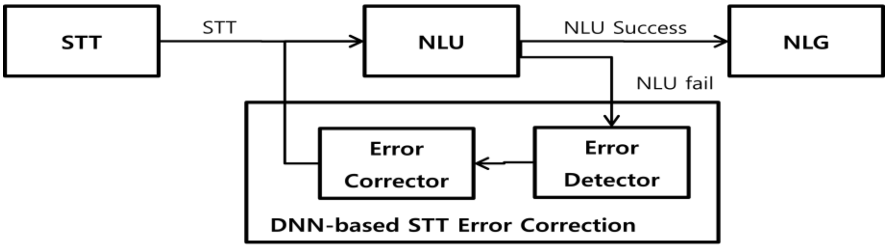
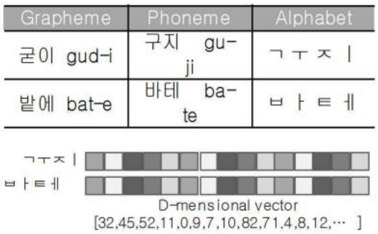


Figure 1. DNN-based STT Error Correction System

The DNN-based STT Error Correction System consists of G2P Convertor, Encoder, Error Detector, and Error Corrector.

- **G2P(Grapheme to Phoneme) Convertor.** This converter converts graphemes of STT data to phonemes. Phonemes converted in the G2P converter can be embedded into vectors for each phoneme unit, as shown in Figure 2 [4].



Grapheme	Phoneme	Alphabet
굳이 gud-i	구지 gu-ji	ㄱ ㅜ ㅈ ㅣ
밭에 bat-e	바테 ba-te	ㅂ ㅏ ㅌ ㅔ

Figure 2. Graphic-to-Vector Embedding

- **Siamese Neural Networks – LSTM Encoder.** This encoder converts graphemes of speech recognition (STT) transcription data into phonemes and embeds the converted phonemes into a vector. By mapping vectors embedded in phoneme units into the vector space for each token, both correct speech tokens and error speech tokens with phoneme similarity are embedded so that they are located in the dense vector space for each correct token. Therefore, the encoder inputs one or more error speech and the corresponding normal speech into each sub-network and learns the parameters within each sub-network so that both the error speech and the normal speech are located in a similar vector space. The construction of this encoder involves inputting one or more error utterances ('트이아스') and the corresponding normal utterance ('트와이스') into each sub-network so that both error utterances and normal utterances are located in a similar vector space. Learn the parameters within. Sub-networks that have completed learning have the same parameters, and ultimately, multiple identical encoder models are built (Figure 3).

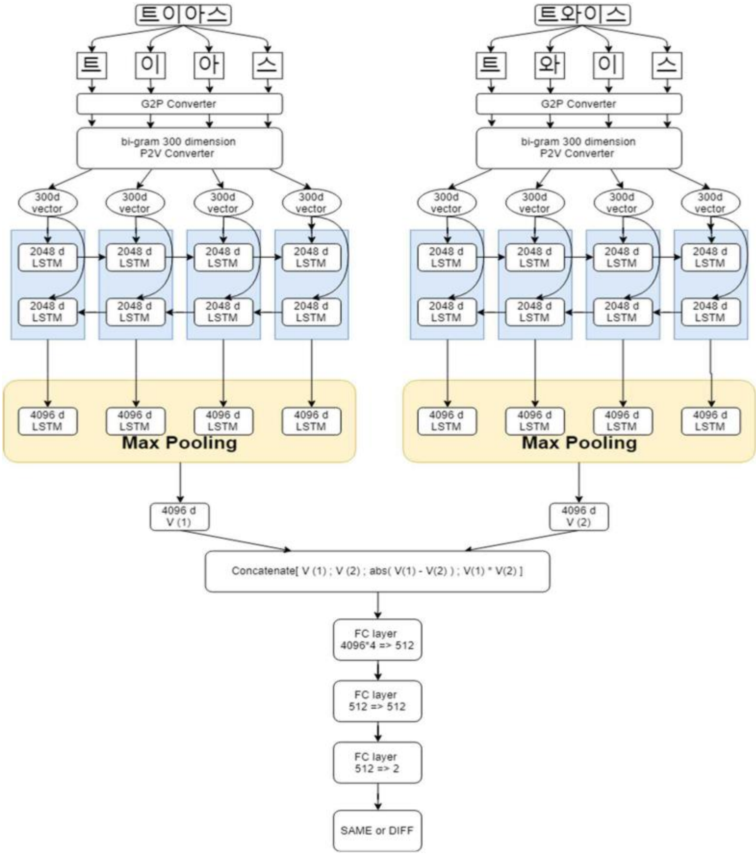


Figure 3. Siamese Neural Networks - LSTM Encoder Architecture

- **Error Detector.** This error detector is intended to detect one or more utterance tokens within a preset Manhattan distance as an error utterance token based on the correct utterance token in the dense vector space for each embedded correct utterance token. This involves learning and classification using encoded token vectors. For example, assuming that the correct utterance token tagged as the correct answer in the dense vector space for each token illustrated in Figure 4 is '방탄소년단', there is at least one correct answer tag without the correct answer within the preset Manhattan distance based on this correct utterance token. The utterance token '달탐사소년단 ' can be detected as an error utterance token. This error detector uses encoded token vectors to perform learning and classification. Based on the correct utterance in the embedding space, utterances within a certain Manhattan distance were designated as error utterances, and an error detector was built based on CNN[5] using the error utterances and correct utterances.

- **Error Corrector.** This error corrector corrects errors by extracting the closest correct speech token based on Manhattan distance as the correct answer from the erroneous speech token detected through the error detector within the dense vector space for each correct answer token embedded through the encoder. For example, if the utterance token '달탐사소년단' is detected as an error utterance token in the dense vector space for each correct token as shown in Figure 4, the tagged object closest to the error token '달탐사소년단' based on the Manhattan distance is The correct utterance token '방탄소년단' is extracted as the correct answer and the error is corrected. The correct utterance through the error corrector is output after intention analysis in NLP. The error corrector determines whether the intention analysis was successful in NLP and, if successful, outputs the corrected correct utterance as is. If the intent analysis for the utterance extracted as the correct answer fails, the error corrector excludes the previously extracted correct answer and extracts the token with the next closest Manhattan distance.

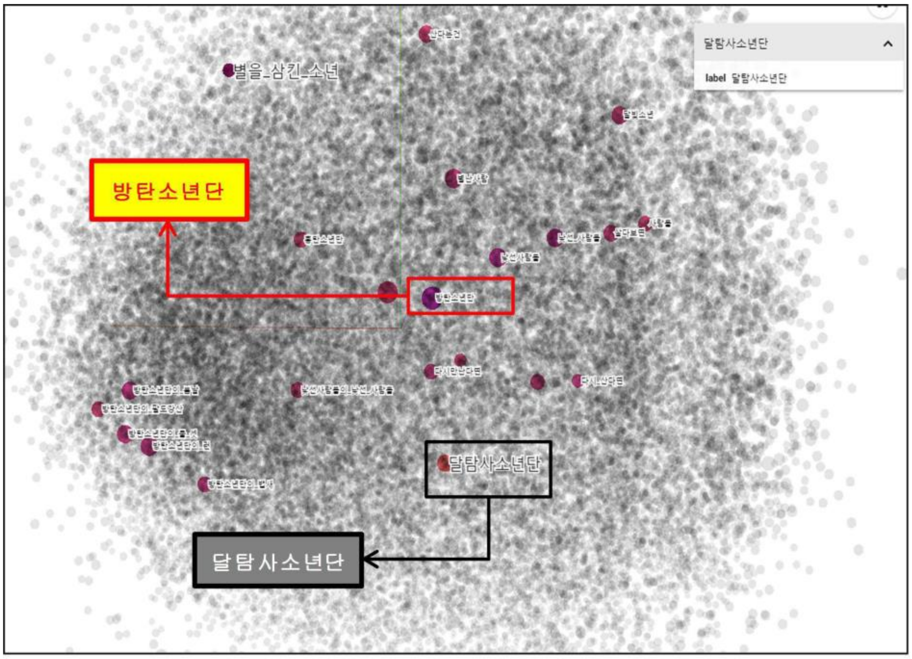


Figure 4. Dense Vector Spatial Pronunciation Similarity Viewer for Each Embedded Correct Answer Token

4. Conclusion

The proposed system in this paper detects and corrects voice recognition errors before natural language processing, which has the effect of improving the success rate of intent analysis in natural language processing without the need to continuously update different versions of voice recognizers and similar word dictionaries in various service domains. In actual commercial voice assistant services, the NLP intent analysis success rate improved from 90.4% to 92.1%. The user utterances used to measure the success rate were 10,000 user utterances from two domains (music streaming, product ordering).

In other words, through this study, the accuracy of intent analysis and service success rate can be increased by additionally correcting speech recognition errors that were not self-corrected in the acoustic model and language model of speech recognition in the pre-intention analysis stage, and new words and product names that are treated as exceptions in the acoustic model. It has the effect of correcting the back.

References

[1] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese Neural Networks for One-shot Image Recognition," Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.
https://www.cs.cmu.edu/~rsalakhu/papers/oneshot1.pdf

[2] H Sak, AW Senior, and F Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," INTERSPEECH (2014), Feb 2014.
https://static.googleusercontent.com/media/research.google.com/ko//pubs/archive/43905.pdf

[3] Chin-Hong Shih,Bi-Cheng Yan, Shih-Hung Liu, and Berlin Chen, "Investigating Siamese LSTM Networks for Text Categorization," Proceedings of APSIPA Annual Summit and Conference 2017.
https://ieeexplore.ieee.org/document/8282104

[4] Kanishka Rao, Fuchun Peng, Haşim Sak, and Françoise Beaufays, "Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),19-24 April 2015.
https://ieeexplore.ieee.org/document/7178767

[5] Yoon Kim, "Convolutional Neural Networks for Sentence Classification," In Conference on Empirical Methods in Natural Language Processing, 2014.
https://arxiv.org/abs/1408.5882

[6] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, "Sequence to Sequence Learning with Neural Networks," NIPS, September 2014.
https://arxiv.org/abs/1409.3215

[7] Heiga Zen and Hasim Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," ICASSP, April 2015.
https://ieeexplore.ieee.org/abstract/document/7178816

[8] Jong-Eon Lee and Boram Lee, Device and Method for Speech recognition error correction. KR Patent 102324829, 2019.