SEQUENCE-TO-SEQUENCE KOREAN PHONEME-TO-TEXT CONVERSION FOR KOREAN SPEECH RECOGNITION

Hosung Park¹, Soonshin Seo², Minkyu Lim³, Donghyun Lee⁴, Yoseb Kang⁵, Juneseok Oh⁶ and Ji-Hwan Kim*⁷

¹³⁴⁵⁶⁷Sogang University, Korea
(E-mail: hosungpark@sogang.ac.kr, lmki@sogang.ac.kr, redizard@sogang.ac.kr, kJoe@sogang.ac.kr, ohjs@sogang.ac.kr, and kimjhwan@sogang.ac.kr)
²Hankuk University of Foreign Studies, Korea
(E-mail: sshy0416@hufs.ac.kr)
*Corresponding author: Ji-Hwan kim

ABSTRACT - In this paper, phoneme-to-text conversion based on sequence-to-sequence learning for speech recognition of Korean standard languages is proposed. To model the difference of Korean phonological change phenomenon, an RNN-LSTM network is used. 22.28% character error rate was achieved in an experiment with 2,200 Korean standard language sentences.

Keywords: Recurrent neural network, Long short-term memory, sequence-to-sequence learning, Grapheme-to-phoneme, Phoneme-to-text

1. INTRODUCTION

Speech recognition outputs the word sequence having the highest probability of the input speech. Recognition consists of four components: acoustic model, language model, decoding network, and lexicon. A phoneme, which indicates the actual pronunciation of the character, is used in converting an audio signal into text. The most important role of the decoding network is to generate an accurate word sequence from a phoneme sequence based on a pronunciation dictionary. Rule-based or statistical automatic phoneme generation is widely used. It is difficult to generate a phoneme sequence based on rules for a foreign language or dialect, as well as convert a phoneme into text. The performance of a statistical method depends on the amount of phonemes and text data. This study proposes Korean phoneme-to-text conversion based on a statistical method.

2. RELATED WORKS

Previous studies of phoneme-to-text conversion have been made to reduce out-of-vocabulary (OOV) words [1]. In this study, the phoneme-to-text conversion without a pronunciation dictionary was performed to remove OOV words. Another study demonstrated the generation of text from vectors extracted directly from an audio signal using connectionist temporal classification [2].

3. PHONEME-TO-TEXT CONVERSION

Generating phoneme sequences based on a statistical method requires a syllable conversion considering preceding and following syllables. A recurrent neural network (RNN) is a representative network considering sequential data [3]. Additionally, a long short-term memory (LSTM) architecture is known to be a good solution of the vanishing gradient problem in a deep structural network like RNN [4]. An RNN-LSTM sequence-to-sequence model is used in this study.
Figure 1 shows the proposed sequence-to-sequence model using an RNN-LSTM. The network consists of an encoder network and a decoder network. The encoder learns input sentences as vectors and the decoder learns the relationship between text and the output of the encoder. If a text sequence is input to the encoder network sequentially, the network learns the relationship between a phoneme and text. The decoder is learned from the last output of the encoder and prints the appropriate text.

4. EXPERIMENT

Grapheme-to-phoneme (G2P) is used to verify the performance of the proposed phoneme-to-text conversion. The phoneme-to-text conversion used in the experiment is modeled using 240 million Korean sentences. Phoneme sequences generated by G2P for these sentences were used as training data and the test set consisted of 2,200 sentences. The experimental result was measured at 22.28% character error rate (CER).

5. CONCLUSION

This paper proposes phoneme-to-text conversion based on sequence-to-sequence learning for speech recognition of Korean standard languages. An RNN-LSTM network was used to model the Korean phonological change phenomenon. Experiment showed 22.28% CER for 2,200 Korean standard language sentences.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2017-0-01772, Development of QA systems for Video Story Understanding to pass the Video Turing Test)

REFERENCES