GRAPHEME-TO-PHONEME CONVERSION SCHEME FOR SENTENCE-BY-SENTENCE LEARNING OF KOREAN MANUSCRIPT USING JOINT SEQUENCE STATISTICAL MODEL

Mousmi A. Chaurasia
Professor IT, MJCET Hyderabad

Abstract

Grapheme-to-Phoneme (G2P) conversion plays an important role in speech synthesis and recognition. In this paper, we used Joint Sequence model where relation between input and output sequence can be generated from a common sequence of joint units. This method is tested on large Korean corpus. The percentage of correct words conversion is 99.94% and 99.99% of word-by-word learning and sentence-by-sentence learning approaches respectively including all exceptional words. Performance of these approaches reduced the relative rate of phoneme error and word error which are relatively higher in other researches.

key words: Grapheme-to-Phoneme(G2P), Joint Sequence model, word-by-word learning approach, sentence-by-sentence learning approach, Korean text.

1. Introduction

Grapheme to Phoneme (G2P) has an inevitable role in natural language processing, speech synthesis as well as spoken dialog systems development. Grapheme to Phoneme conversion is the process to produce a pronunciation for an input word (spelling only) to build a pronunciation dictionary. These conversions are based on some pre-defined rules which are usually created by a automated statistical analysis of a pronunciation dictionary. The main purpose to build pronunciation dictionaries is to provide pronunciation specifically to those words which do not appear in the dictionary.

In the year 2008, Bisani and Ney proposed a joint sequence model based on co-sequence and grapheme-phoneme joint multigram. They concurrently aligned and segmented a pronunciation dictionary using discounted Expectation Maximization (EM) scheme and joint sequence modeling into a successful G2P model. The fundamental idea of joint-sequence models is to generate the relation of input and output sequences from a common sequence of joint units which carry both input and output symbols. In the simplest case, each unit carries zero or one as input and zero or one as output symbol. It was a rational approach to grapheme-to-phoneme conversion which was profounder on statistical decision theory. They experimented this approach on English, German and French with promising results and minimal phoneme error rate (PER) and word error rate (WER) [1]. This approach currently sets the prominent and precise standard for G2P conversion. This model[1] we have used in our experimental purpose for training and testing of Korean corpus. We assessed this model in two kinds of approaches:- word-by-word learning and sentence-by-sentence learning approach for Korean G2P conversion. Section 2 describes the methodology of Korean G2P converter. Elaborate explanation on results and discussion has provided in Section 3. Lastly, conclusion and future work has been drawn.

2. Korean Orthography and Methodology used in G2P conversion

In korean writing, a set of jamo(alphabets) is collection of an eumjeol (syllables), and a series of eumjeol is grouped into eojeol (words). An eumjeol consists of up to three components: an initial consonant (onset), a vowel (nucleus), and a final consonant (coda). Thus, each eumjeol can be separated into three parts like onset-nucleus-coda(CVC). Korean alphabets set consist of 21 vowels and 19 consonants.

Romanization is the translation of sounds of a foreign language into English letters. Romanization of Korean words allows those who can't read Korean to phonetically pronounce it. From [2] , Korean segments using jamo(Hangul alphabets) and romanization are being referred. We have used statistical approach for G2P analysis. G2P converter takes Korean texts in English transliteration as input. From here, our approach is divided into two fields:

1. Word-by-word learning approach
2. Sentence-by-sentence learning approach

In both approaches, removes all punctuation marks except '-', '@', '#', '$', '%' , '^' , '&' , '*' , '_' , '!'. The chances of error to extract grapheme from raw data are almost negligible because we used 1-to-1 alignment from Korean alphabets to transliteration in English. For both the learning approach, transliterated output send as input to G2P converter. Since this is statistical approach, so each grapheme has predefined rules to generate the phonetic transcription of the current grapheme. Once the grapheme is produced, it will generate the correct phonetic transcription of original Korean sentences. For word-by-word learning approach, converter tokenized the words into graphemes. Furthermore, graphemes are mapped to their corresponding phonemes. The pronunciation of each grapheme may vary, depends on the occurrence of the preceding and succeeding graphemes. So during alignment of each grapheme to its respective phoneme, system needs to check the preceding and succeeding graphemes.

Previous researches was relied on word learning approach which commonly lengths the duration of grapheme conversion. Nevertheless, conversion of grapheme to phoneme are yet to be captured before concluding the results extracted from testing data. It takes several days in constructing the training models and testing the test data. Sentence-by-sentence learning approach is the solution for word tokenization problem . It has unlimited context length feature. It is not limited to number of words or context length per line. This practical solution reduces the processing time of training and testing data. It takes only few hours to supply the final results.

3. Results and Discussion

The corpus used in the experiment was the transcription of conversational style speech synthesis database provided by ETRI (Electronics and Telecommunications Research Institute). It was a recording of simulated conversational sentences performed by a professional female announcer. The pronunciation was manually transcribed by a native Korean speaker. Originally, corpus contains 18,030 sentences . For the experiment, the corpus should divided into two modes:

1. 80-20 approach: Whole corpus was divided into two portions : one was training data (80%; 1,06,217 words or 14424 sentences) and other was evaluation test data (20%; 20,069 words or 3606 sentences).

2. 5fold cross validation: We conducted K-fold-cross-validation in the experiments. In this research, K = 5 was adopted. The data were randomly partitioned into five equal folds, each including 25260words or 3606 sentences from whole document. Whole document division based on words was target as input text for word-by-word learning approach and partition based on sentences was aimed as input data for sentence-by-sentence learning approach. On each round of experiment, four folds were used a training data set, and the remaining fold was used as the testing data set. Since a jackknife approach was used, each fold was in a test set once and in a training set four times.

We conducted each round of experiment by increasing the number of training examples with each experiment; this was repeated 5-fold. Finally, the five results from the folds will be averaged to produce a single estimation. All the reported results are on the evaluation test data. Four metrics were used to evaluate performances:

(1) Phoneme level accuracy (a measure which scores insertion, deletion, and substitution errors equally),

(2) Phoneme error rate (PER) (an edit distance between the automatic transcription result (candidate) and reference pronunciation divided by the number of phonemes in the reference pronunciation.)

(3) Word level accuracy (a measure which counts a word as correct if all the phonemes within it are correct).

(4) Word error rate (WER) (the relative proportion of words that have at least one phoneme error)

Table1 and Table2 present the performance of G2P conversion accuracy and error rate of 80-20 approach and 5fold cross validation approach of word-by-word learning and sentence-by-sentence learning approaches respectively.

TABLE 1 G2P conversion accuracy our proposed approach

Joint Sequence Model [1]	Word-by-word approach-- Testing data	
	80-20 approach	5fold approach
Phoneme level accuracy	99.97	99.99
Word level accuracy	99.70	99.94
Phoneme Error Rate (PER)	0.022	.07
Word Error Rate (WER)	0.294	.07

TABLE 2 G2P conversion accuracy our proposed approach

Joint Sequence Model [1]	Sentence-by-sentence approach-- Testing data	
	80-20 approach	5fold approach
Phoneme level accuracy	99.99%	99.99%
Word level accuracy	99.97%	99.99%
Phoneme Error Rate (PER)	.002%	.0004%
Word Error Rate (WER)	.024%	.006%

TABLE 3 Comparison of G2P conversion accuracy of 80-20 approach with previous related research of word-based approach

	Testing data	
	Phoneme level (%)	Word level (%)
Our Approach	99.97	99.70
[2]	99.216	94.105
[3]	99.057	92.927
[4]	98.354	86.629

Table3 shows the comparison among previous state-of-the-art related work performances tested on same Korean database. Lee and Lee [2] had proposed rule based system for Korean G2P TTS. They experimented the corpus on rule-based data-driven model. They measured their performances based on phoneme level and word level accuracy with full rules and rule pruning procedures. They obtained 99.22 percent phoneme-level accuracy with full rules and only a small performance degradation (0.1percent) by using reduced rule approach. With reduced rules, they obtained 99.11 percent phoneme level accuracy [2]. The downside part of rule-based data-driven model[2] was that it had limited context length feature which varies up to maximum ten. However, Lee et.al[2] had compared other data driven methods like decision trees [3] and Hidden Markov model (HMM) [4] on the same Korean corpus. They applied graphemes contexts (five letters both left and right side) for decision trees and 5-gram grammar for HMM based methods. The final consequences of Korean corpus based on word-level accuracy of decision trees and HMM models are 92.927 percent and 86.629 percent respectively. It is evident from results that joint sequence model[1] performs much better than rule-based data-driven model[2] better than decision trees algorithm better than HMM. The results illustrate that our method surpassed other data-driven methods on the same corpus.

In this paper, we discussed an competent and robust statistical approach for G2P conversion for Korean language. The system has been implemented in python. Our Sentence-by-sentence learning approach was not restricted to length of sentence. This approach has ability to learn a sentence with endless words in single line. This is most profitable fixation of grapheme to phoneme conversion. Moreover, 1-to-1 alignment scheme has been opted for alignment of each grapheme to phoneme conversion. Therefore, it reduced the error prospect and increases the precision of results. In English, there are 1-to-n and n-to-1 alignment as well as 1-to-1 alignment. Moreover, since the Korean writing system is more or less phonetic, most graphemes are enunciated as they are referred from pronunciation-varying verbalization database. At the last, the size of Korean database have an adequate amount of words which cover most of the pronunciation changes. This approach can be used in various speech recognition and speech synthesis application.

Conclusion and Future work

This paper introduced a sentence-by-sentence learning approach for G2P conversion. Aforementioned approach reduced the processing time of building the models from training corpus and fetching results from testing data. During experimentation, we had also tested our corpus on word-by-word learning approach and from the end result of both approaches, it is explicitly suggested that sentence learning furnished more precise outcome than word learning approach for G2P conversion. The noteworthy parameter was phoneme error rate (PER) and word error rate (WER) , which were almost negligible. This concluded our paper that with high transcription rate of G2P, our sentence-by-sentence learning approach was more efficient, methodical and robust.

Acknowledgement

We would like to thank Sogang University, Seoul for providing Korean corpus used in this exp.

References

[1] M.Bisani, H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion", Speech communication vol.50, pp 434-451, January 2009.

[2] J. Lee, G.G. Lee, " A data-driven grapheme-to-phoneme conversion method using dynamic converting rules for Korean TTS systems", Computer Speech and Language vol.23, no.4, pp. 423-434, January 2009.

[3] A. Black, K. Lenzo, V. Pagel, "Issues in building general letter to sound rules", Proceeding of Third ESCA Workshop on Speech synthesis, pp. 77-80, 1998.

[4] P.Taylor, "Hidden Markov model for grapheme to phoneme converison", Proceeding of Interspeech, Libson, pp. 1973-1976, 2005.