AN ERROR ANALYSIS OF SEQUENCE-TO-SEQUENCE NEURAL NETWORKS ON ENGLISH PHONETIC ALPHABET CONVERSION

Kong Joo Lee
Dept. of Radio and Information Communications Engineering, Chungnam National University, Korea
kjoolee@cnu.ac.kr

Jee Eun Kim
Dept. of English Linguistics, Hankuk University of Foreign Studies, Korea
jeeeunk@hufs.ac.kr

Abstract: English words present inconsistency between their spelling and pronunciation, which requires phonetic alphabet for accurate pronunciation. However, each English dictionary not only lists a different set of pronunciation from one another, but also adopts a phonetic alphabet represented with different notation. These differences in pronunciation and phonetic notation confuse English language learners. A recent research on automatic conversion of different pronunciations shows a result with the accuracy between 74.5 ~ 89.6% produced utilizing sequence-to-sequence (seq2seq) model, a popular mechanism used in deep-learning.
This research suggests an error analysis conducted on the results of automatic conversion of different pronunciations. The errors are bidimensional. One dimension is classified into types of segment and suprasegment: consonant, vowel, stress errors. The other describes the types of errors including addition, deletion and alternation. The purposes of the error analysis can be summarized as follows: 1) to survey different phonetic alphabet systems for English and figure out the characteristics of each system, 2) to verify various pronunciation rules and the context information identified in automatically converted data, and finally 3) to suggest a guideline for organizing a training set to be used for learning the seq2seq model in automatic pronunciation conversion.

Keywords: English pronunciation, phonetic alphabet, sequence-to-sequence (seq2seq) model, error analysis

Introduction

The English spelling system does not represent the pronunciations of words: a written word does not necessarily express its vocal sound (Jurafsky, 2000). A phoneme can be expressed with many different letters of the alphabet, which causes difficulties in predicting an accurate pronunciation. Moreover, some letters are not pronounced at all, and mapping a phoneme to a letter can be represented utilizing one-to-many or many-to-one relationship. In other words, each phoneme is represented by more than one written letter or a sequence of letters, and a letter can be mapped to more than one sound. These types of inconsistency between the sound of a word and its spelling make learning English pronunciation difficult. Providing a phonetic alphabet system in a dictionary may help the learners of English to master the pronunciation. Many English dictionaries, however, present a system of their own, in which different phonetic alphabets are used to represent the same sound. Another inconsistency in phonetic alphabets also confuses the learners and becomes a cause to produce pronunciation errors.

This research aims to analyze the errors produced while automatically converting the phonetic alphabets to the notations adopted in other dictionaries (Lee and Choi, 2017). Those dictionaries include four different online English dictionaries: CMU Pronouncing Dictionary (CMUdict), New Oxford American Dictionary (NOAD), Merriam-Webster Collegiate Dictionary (MWCD), and New Oxford Dictionary of English (NODE)[1]. All the dictionaries provide phonetic alphabets to represent the pronunciations of each entry, but the notations are all varied from one another. Each of the four sets of phonetic alphabets is automatically converted to the rest three notations. The conversion was performed using a sequence2sequence model which outperforms other models when the length of the input sequence is not identical to that of the output. The average accuracy of the conversion

[1] CMUdict: http://www.speech.cs.cmu.edu/cgi-bin/cmudict
NOAD: https://en.oxforddictionaries.com/definition/us/
MWCD : https://www.merriam-webster.com/
NODE: https://en.oxforddictionaries.com/

is 83.1% and some of the conversion between particular sets exceeds 89%. When the conversion process has completed, conversion errors were collected for an analysis by which various types of errors were identified. This paper, however, focuses on the consonant errors, more specifically those which identified the errors between American and British English in particular. These types of errors are found when the conversion was processed between NOAD and NODE.

Characteristics of Phonetic Alphabets and Dictionaries

This research has initially selected four different online dictionaries: CMUdict, NOAD, MWCD and NODE. CMUdict is developed as an online dictionary for speech processing and has a reputation for its pronunciation representations. MWCD and NODE are traditionally well-known since their paper versions have been widely used. NOAD is selected because it is the American English counterpart of NODE. The pronunciations of these lexicons are transcribed using four different phonetic alphabet systems. CMUdict uses ARPAbet and NODE has adopted International Phonetic Alphabet (IPA). NOAD and MWCD have utilized the systems of their own named as NOAD and MWCD respectively.

CMUdict is originally created for use in speech recognition and lists the most number of entries, 117,413 which includes inflected forms. It transcribes North American pronunciations with utilizing 39 phonemes composed of 23 consonants and 16 vowels. NOAD is constructed based on NODE, but its 24,727 entries present the pronunciations of contemporary American English, represented with 42 phonemes among which there are 24 consonants and 19 vowels. MWCD contains 24,988 entries with American English pronunciations which are transcribed with 45 phonemes, 25 consonants and 20 vowels. NODE includes 24,458 entries whose pronunciations mainly represent British English. The pronunciations are transcribed with the most number of phonemes, 49 which are comprised of 26 consonants and 23 vowels. In addition, all the dictionaries present pronunciation variations for an entry if there is any. The average number of variations ranges from 1.069 to 1.313. However, the variations are often specified with partial transcriptions, providing only the different part of the full transcription. A dictionary entry is listed with its full pronunciation and optionally with the varied phoneme or the syllable(s) which includes the variations.

Automatic Conversion of Phonetic Alphabet for English Using Seq2seq Model

Deep Neural Networks (DNNs) have recently accomplished the state-of-the-art performance on various pattern-recognition tasks including speech and vision. They are remarkably powerful and flexible in resolving complicated problems for not only vision, but also language/speech processing such as machine translation in particular. Among language/speech processing applications adopting DNNs, Recurrent Neural Networks (RNNs) are a popular class since they can process arbitrary sequences of inputs. However, they show a limitation that they work with vectors of fixed dimensionality. In other words, they can achieve the expected performance only when the lengths of source and target are identical while some of language/speech related problems are represented with the sequences whose length is subject to change. Because of the limitation of RNNs, this research adopted a seq2seq model which consists of two RNNs utilizing sequential information.

Seq2seq Model

In the seq2seq model, one RNN computes a sequence of input as an encoder and the other generates the output as a decoder. The encoder RNN maps the various lengths of the input to a fixed-sized vector, and then the decoder RNN generates various lengths of output sequences from the vector (Cho et al., 2014).

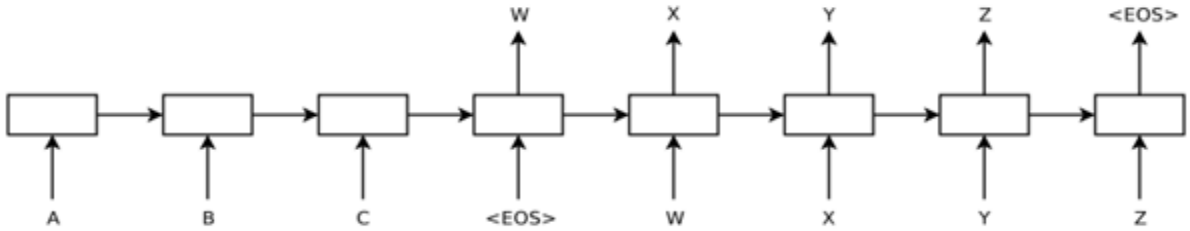

Figure 1. Basic Architecture of a seq2seq model (Sutskever et al., 2014)

The basic architecture depicted in Figure 1 represents the cells of the RNN, by which a sequence is processed. An encoder computes the input, ‘ABC’, creates a fixed-sized vector from which a decoder produces the output

sequence 'WXYZ' stopping when it reaches the end-of-sentence, <EOS>. Each sentence always ends with a special symbol, usually specified with 'EOS' by which all the possible lengths of sequences can be defined in the model. In a language/speech processing system utilizing the model, the probability of 'WXYZ<EOS>' is computed according to the representation of 'ABC' which is also computed by the first RNN.

This research has implemented a seq2seq model on automatic conversion of phonetic alphabets for English. The model is complemented to process different lengths of input and output sequences. As described with Figure 1 above, the model has successfully generated an output sequence whose length is different from that of the input.

Normalization of Dictionary Entries

The pronunciation transcriptions of all the entries were extracted from the four dictionaries. Since those pronunciation transcriptions utilizing phonetic alphabets are varied for the same entries, all of the transcriptions are listed under each entry. If there exist pronunciation variations, they are also listed after they are restored to a full pronunciation sequence.

Table 1: Examples of Pronunciation of MWCD

Entry	Pronunciations before Conversion	Pronunciations after Conversion
acceleration	/ik-ˌse-lə-ˈrā-shən, (ˌ)ak-/	ik-ˌse-lə-ˈrā-shən ˌak-ˌse-lə-ˈrā-shən ak-ˌse-lə-ˈrā-shən
proselytization	/ˌprä-s(ə-)lə-tə-ˈzā-shən, ˌprä-sə-ˌlī-tə-/	ˌprä-slə-tə-ˈzā-shən ˌprä-sə-lə-tə-ˈzā-shən
conference	/ˈkän-f(ə-)rən(t)s, -fərn(t)s/	ˈkän-fə-rənts ˈkän-fə-rəns ˈkän-frənts ˈkän-frəns

The variations in all the dictionaries are presented as shown in the second column of Table 1 under "Pronunciation before Conversion". For example, two pronunciations are mapped to an entry '*conference*'; 1) /ˈkän-f(ə-)rən(t)s/ and 2) / -fərn(t)s/. The first transcription represents a full pronunciation although it includes two options by which a weak vowel known as schwa can be reduced and the voiceless alveolar stop, /t/ can also be omitted. The third column under "Pronunciations after Conversion" displays all the variations restored as full pronunciations which are generated calculating the possible combinations with the given options. Restoring the variations to the corresponding full pronunciation is performed as preprocessing. However, some of the restored sequences are eliminated from the final output list if there is a difference in the total number of syllable[2].

When all the variations are extended to a full pronunciation, an appropriate stress pattern is added to complete the pronunciation. Lexical stress is relative emphasis to a certain syllable, which is caused by phonetic properties such as loudness or vowel length. English has one or more syllables in every word and stress is placed on one or more syllables. Lexical stress in English presents different levels including primary, secondary and tertiary although the position of each level is less predictable. Accordingly, English stress patterns have to be memorized as part of the pronunciation. More importantly, lexical stress in English is contrastive to distinguish parts of speech. Given a pair of identical strings, they convey two different meanings from each other if stress is placed on different syllables. Some words may contain more than one stressed vowel, but exactly one of the stressed vowels is more prominent than the others. Each dictionary presents a unique system to indicate the stress pattern of a word and shows the different number of stress for the same word. Moreover, British and American English have different stress patterns from each other, which also leads to reduction of a vowel. These differences result in dissimilar pronunciations.

[2] This process will be complemented in the next step of the research so that all the restored sequences can make the final output list.

Table 2: Stress Systems

Dictionary	*insignia*	*rare*	*rehab*
CMUdict	IH2 N S IH1 G N IY2 AH0	R EH1 R	R IY0 HH AE0 B
NOAD	inˈsignēə	rer	ˈrēˌhab
MWCD	in-ˈsig-nē-ə	ˈrer	ˈrē-ˌhab
NODE	ɪnˈsɪgnɪə	rɛː	ˈriːhab

Table 2 displays the differences in the stress pattern of the dictionaries. The stress is placed before the stressed syllable in NOAD/MWCD/NODE whereas it is indicated after the stressed vowel in CMUdict, using a number, 0~2. There is no special indication in IPA/MWCD/NOAD for no stress while '0' is inserted in CMUdict.

Table 3: Stress Description

Phonetic Alphabet	**Primary Stress**	**Secondary Stress**	**No Stress**
IPA / MWCD / NOAD	[']	[,]	
ARPAbet	1	2	0
Standardization	*	%	

Different stress systems are standardized in order to train automatic conversion of pronunciations among the dictionaries. Table 3 shows the standardized notation of stress patterns. The primary stress is represented with the asterisk symbol '*' while the secondary is indicated using the percent symbol, '%'. Both of them are inserted at the end of a stressed syllable.

When adding a stress pattern to all the entries is completed, non-ASCII codes are mapped to ASCII ones. NOAD, MWCD and NODE contain some of the non-ASCII code to represent phonemes. The mapping process is necessary in order to ease the conversion process. In addition, the dash symbol, '-' is inserted between the phonemes for clear distinction.

Table 4: Sample Input Words for Seq2seq Model

		insignia	*rare*	*rehab*
Original Entries	CMUdict	IH2 N S IH1 G N IY2 AH0	R EH1 R	R IY0 HH AE0 B
	NOAD	inˈsignēə	rer	ˈrēˌhab
	MWCD	in-ˈsig-nē-ə	ˈrer	ˈrē-ˌhab
	NODE	ɪnˈsɪgnɪə	rɛː	ˈriːhab
Normalized Entries	CMUdict	IH-%-N-S-IH-*-G-N-IY-%-AH	R-EH-*-R	R-IY-HH-AE-B
	NOAD	I1-N-S-I1-*-G-N-E2-E3	R-E1-R	R-E2-*-H-A1-%-B
	MWCD	I1-N-S-I1-*-G-N-E2-E3	R-E1-*-R	R-E2-*-H-A1-%-B
	NODE	I2-N-S-I2-*-G-N-I2-E3	R-E2:	R-I1:-*-H-A4-B

Table 4 lists normalized word examples to be used as input for training which utilizes a seq2seq model. The stress pattern is now represented with the standardized notation for the entries in all the dictionaries.

Automatic Conversion with Seq2seg Model and Experimental Results

To train the seq2seq model, 12 models have been constructed since the pronunciations of each dictionary have to be converted to those transcribed in the rest of the dictionaries. Each dictionary is coupled with the rest of the dictionaries, which results in 6 conversion pairs prepared for bidirectional conversion.

Table 5: Size of Training Data

Conversion Pairs	Number of Entries in Training Data
MWCD(MWCD) ↔ NODE(IPA)	44,086
MWCD(MWCD) ↔ CMUdict(ARPAbet)	37,471
MWCD(MWCD) ↔ NOAD(NOAD)	45,549
CMUdict(ARPAbet) ↔ NODE(IPA)	34,664
CMUdict(ARPAbet) ↔ NOAD(NOAD)	35,082
NODE(IPA) ↔ NOAD(NOAD)	44,403

With the normalized entries completed, training data sets are created for all the six pairs. The size of the training data varies for each pair as presented in Table 5. Although each model requires a different setting of parameters for training, all the models have been trained using the optimized parameters set for the two models. When the training is completed, a test data set is created with 1,000 entries, all of which are commonly listed in the four dictionaries. With this testing data, the conversion performances are evaluated.

Table 6: Accuracy of Conversion Results

Conversion	Accuracy	Conversion	Accuracy
NOAD(NOAD) → MWCD(MWCD)	89.6	MWCD(MWCD) → CMUdict(ARPAbet)	82.0
CMUdict(ARPAbet) → MWCD(MWCD)	88.9	NOAD(NOAD) → CMUdict(ARPAbet)	81.5
MWCD(MWCD) → NOAD(NOAD)	87.8	MWCD(MWCD) → NODE(IPA)	81.3
CMUdict(ARPAbet) → NOAD(NOAD)	87.6	CMUdict(ARPAbet) → NODE(IPA)	79.7
NODE(IPA) → MWCD(MWCD)	83.8	NOAD(NOAD) → NODE(IPA)	77.9
NODE(IPA) → NOAD(NOAD)	82.1	NODE(IPA) → CMUdict(ARPAbet)	74.5

Table 6 presents the accuracy of automatic conversion performed on each pair of the models. The highest accuracy is 89.6% produced when NOAD(NOAD) is converted to MWCD(MWCD). The result seems natural since both dictionaries transcribe the pronunciation of American English. On the other hand, the conversion from NODE(IPA) to CMUdict(ARPAbet) shows the lowest performance resulting in the accuracy of 74.5%. The low performance seems attributed to the fact that NODE transcribes British English whereas CMUdict provides the pronunciation of American English. This can be confirmed by the other two results with low accuracy produced by the conversion from NOAD(NOAD) to NODE(IPA) and from CMUdict(ARPAbet) to NODE(IPA). All of the three models convert British English from/to American English.

Error Analysis of Automatic Conversion

This research focuses on the results produced while converting NOAD to NODE and vice versa. Since they transcribe the pronunciation of American and British English respectively, the results are considered as useful sources to detect the meaningful conversion errors. The conversion result is analyzed and classified according to the type of errors. Three prominent types include stress, vowel, and consonant. This research, however, focuses on consonant related errors. American and British English present differences in a set of consonants transcribing the same word.

Table 7: Pronunciation Variations of Consonants

British English	American English
(❒)	❒
♦	☼
♦	☾
⌘	♦
♦	♦
❄	❄

Table 7 lists up the pronunciation variations in some of the consonants between British and American English (Hosseinzadeh, et al., 2015). One of the most prominent differences is the rhotic /r/. In American English, the letter 'r' is mostly pronounced except the cases in some dialects. However, it is not pronounced in British English when it occurs in the coda position. For example, a word, *park* is pronounced as /park/ in American English while it is dropped in British English as in /pak/.

Table 8: Errors on Converting /r/ #1

		To **NOAD**	
			/r/
From **NODE**	**No /r/** *in* **Testing Data**	Correct: 441 Incorrect: 20	Correct: 191 Incorrect: 23
	No /r/ *in* **Training Data**	None	7,165

The process of converting NODE to NOAD has produced 441 correct cases and 20 incorrect cases of not inserting /r/ as Table 8 has presented. The conversion is incorrect because the rhotic /r/ in British English has to be realized as /r/ in American English, but it failed as in /ˌaftəˈnoͯon/ which is meant to be /ˌaftərˈnoͯon/. Similarly, 23 errors have been detected when /r/ should not be inserted in NOAD since /r/ is neither rhotic nor included in the words listed in NODE.

Table 9: Errors on Converting /r/ #2

Entry	NODE (IPA)	NOAD (NOAD) reference	NOAD (NOAD) model	Description
yoga	ˈjəʊgə	ˈyōgə	ˈyōgər	/r/ insertion error
sigma	ˈsɪgmə	ˈsigmə	ˈsigmər	/r/ insertion error
forerunner	ˈfɔːrʌnə	ˈfôrˌrənər	ˈfôˌrənər	/r/ deletion error
joker	ˈdʒəʊkə	ˈjōkər	ˈjōkə	/r/ deletion error
altogether	ɔːltəˈgɛðə	ˌôltəˈgeT͟Hər	ˌôltəˈgeT͟Hər	correct (/r/ insertion)
benefactor	ˈbɛnɪfaktə	ˈbenəˌfaktər	ˈbenəˌfaktər	correct (/r/ insertion)

Table 9 displays more examples of the results produced by converting /r/. For example, a word, *yoga* should be converted to /ˈyōgə/ rather than /ˈyōgər/ in which the inserted /r/ causes an error. In the training set, there are 7,165 cases in which British rhotic /r/ is correctly restored in NOAD. Table 8 also presents the correctly converted examples such as *altogether*: the last phoneme /r/ in NOAD has been successfully restored.

Another prominent pronunciation variation is mapping /t/ in British English and /d/[3] in American English. The voiceless alveolar stop, /t/ in British English often pronounced as a voiced alveolar flap in American English usually when it occurs in an intervocalic position or the phoneme is represented by a double 't' in its spelling.

[3] This is supposed to be the voiced alveolar flap /☼/, but NOAD uses /d/ to represent it.

Table 10: Errors on Converting /t/ #1

		To **NOAD**	
		/t/	**/d/**
From **NODE**	**/t/** *in* **Testing Data**	Correct: 213 Incorrect: 2	Correct: 55 Incorrect: 4
	/t/ *in* **Training Data**	12,292	3,246

As the result is shown in Table 10, the conversion of /t/ has generated only a couple of errors; British /t/ is supposed to be realized as /d/, but /t/ is produced instead. Another small number of errors are produced when /t/ is converted to /d/ rather than /t/ which is the intended result.

Table 11: Errors on Converting /t/ #2

Entry	**NODE (IPA)**	**NOAD (NOAD) reference**	**NOAD (NOAD) model**	**Description**
prophetess	ˌprɒfɪˈtɛs	ˈpräfədəs	ˈpräfətəs	/d/→/t/ error
ghettoize	ˈgɛtəʊʌɪz	ˈgedōˌīz	ˈgetōˌīz	/d/→/t/ error
aglitter	əˈglɪtə	əˈglɪtə	əˈglɪdə	/t/→/d/ error
footlights	ˈfʊtlʌɪts	ˈfo͡otˌlīts	ˈfo͡odˌlīts	/t/→/d/ error
bottom	ˈbɒtəm	ˈbädəm	ˈbädəm	correct (/t/→/d/)
waiter	ˈweɪtə	ˈwādər	ˈwādər	correct (/t/→/d/)

Table 11 presents a set of examples in which both correct and incorrect conversion examples are listed. The voiceless alveolar stop is trained to be converted to the flap, but it stays in the pronunciation of a word such as *prophetess* of NOAD causing an error. On the other hand, /t/ is converted to /d/ in *footlights* when it should not.

Table 12: Errors on Converting /t/ #3

		To **NOAD**	
		/t/	
From **NODE**	**/t/** *in* **Testing Data**	Correct: 213 Incorrect: None	Correct: None Incorrect: 1
	/t/ *in* **Training Data**	12,292	717

Table 12 shows an interesting result of converting /t/. NODE /t/ has been successfully converted to /t/ in NOAD, without generating an error. When the British /t/ is converted to null in NOAD, only a single error has produced. However, no correct conversion has occurred either.

Table 13: Errors on Converting /t/ #4

Entry	NODE (IPA)	NOAD (NOAD) reference	NOAD (NOAD) model	Description
antler	ˈantlə	ˈantlər	ˈanlər	/t/ deletion error
costly	ˈkɒstli	ˈkôslē	ˈkôslē	correct (/t/ deletion)
frantic	ˈfrantɪk	ˈfranik	ˈfranik	correct (/t/ deletion)
hunter	ˈhʌntə	ˈhənər	ˈhənər	correct (/t/ deletion)

More examples are provided in Table 13. It lists the single error on converting /t/ to null as in *antler* in which /t/ is supposed to remain as is. In many other cases, the voiceless alveolar stop in British English becomes silent in some American English words as in *hunter*, which is not an error. This type of examples is included in the training data, but no such case has been found in the testing data.

Table 14: Errors on Converting /☠/ #1

		To **NOAD**	
		/NG/	**/n/**
From **NODE**	**/☠/** *in* **Testing Data**	Correct: 11 Incorrect: 1	Correct: None Incorrect: 1
	/☠/ *in* **Training Data**	2,028	49

The next case is converting the velar nasal stop /☠/ in British English to the alveolar nasal /n/ in American English. As Table 14 suggests, this is a rather rare case. Although the training data set contains 2,028 correct conversions of /☠/ to /NG/, only 11 correct conversions are identified. Converting /☠/ to /n/ has been trained with 49 conversion cases, but no such cases have been found in the testing process.

Table 15: Errors on Converting /☠/ #2

Entry	NODE (IPA)	NOAD (NOAD) reference	NOAD (NOAD) model	Description
conclave	ˈkɒŋkleɪv	ˈkän¸klāv	ˈkä NG¸klāv	/n/→/NG/ error
congresswoman	ˈkɒŋgrɛs¸wʊmən	ˈkäNGgrəs¸woomən	ˈkängrəs¸woomən	/NG/→/n/ error
exceedingly	ɪkˈsiːdɪŋl	ikˈsēdiNGlē	ikˈsēdiNGlē	correct (/ŋ/→/NG/)
spanking	ˈspaŋkɪŋ	ˈspaNGkiNG	ˈspaNGkiNG	correct (/ŋ/→/NG/)
concrete	ˈkɒŋkriːt	ˈkän¸krēt	ˈkän¸krēt	correct (/ŋ/→/n/)
increase	ˈɪŋkriːs	inˈkrēs	inˈkrēs	correct (/ŋ/→/n/)

Table 15 includes a pair of errors where /☠/ is incorrectly realized as /NG/ instead of /n/ as in *conclave*, and as /n/ when it is intended for /NG/ as in *congresswoman*. It also lists several examples of correctly converted /☠/ to /NG/ such as *exceedingly* and to /n/ as in *increase*.

Table 16: Errors on Converting /p/ #1

		To **NOAD**	
		/p/	**␣**
From **NODE**	**No /p/** *in* **Testing Data**	Correct: None Incorrect: 1	Correct: None Incorrect: None
	No /p/ *in* **Training Data**	64	None

In American English, the voiceless bilabial stop /p/ is mostly pronounced when it appears in its spelling as in *assumption*. However, it is silent in British English. The training data set lists 64 cases of correct conversion in

which the silent /p/ of NODE has been restored to /p/. However, the testing set does not include a single case of the conversion as shown in Table 16.

Table 17. Error on Converting /p/ #2

Entry	NODE (IPA)	NOAD (NOAD) reference	NOAD (NOAD) model	Description
sometime	ˈsʌmtʌɪm	ˈsəmˌtīm	ˈsəmˌptīm	/p/ insertion error

Table 17 presents an example of incorrectly converted /p/ as in an entry, *sometime*. The voiceless bilabial stop is inserted in NOAD when it should not.

Table 18. Examples of Converting /p/ in Training Data

Entry	NODE (IPA)		NOAD (NOAD)	
assumption	əˈsʌmʃn	əˈsʌmʃən	əˈsəmSHn	əˈsəmSHən
	əˈsʌmpʃn	əˈsʌmpʃən	əˈsəmpSHn	əˈsəmpSHən
humpback	ˈhʌmbak	ˈhʌmpbak	ˈhəmpˌbak	
symptom	ˈsɪmtəm	ˈsɪmptəm	ˈsimtəm	ˈsimptəm

Table 18 lists up the cases of correct conversion of /p/ included in the training data. With examining the training data, the cause for the error in *sometime* is detected as shown in Table 17. For example, when an entry has pronunciation variations as in *assumption*, it generated 16 possible cases for the training. Some of the data such as /əˈsʌmʃn/ and əˈsəmpSHən/ has causes the error in *sometime*; even when there is no /p/ in the source data, /p/ is inserted in the target data. In other words, the accuracy of the seq2seq model decreases when there are more than one training data for an input sequence.

Conclusion

This research has presented automatic conversions of phonetic alphabets for English utilizing a seq2seq model. Examining the conversion results helps verifying the differences in pronunciations between British and American English. One of the most noticeable distinctions is pronouncing /r/ and /t/ which represent the typical pronunciations of the two dialects of English.

A seq2seq model was implemented for this research since it is known for its good performance in successfully generating an output sequence whose length is different from that of the input. However, some of the errors suggest that the quality of the training data determines the accuracy of the performance. The model has performed well on the input which a single output is mapped to, but it requires further adjustment on creating a training data set for the cases in which multiple outputs are expected. In other words, high performance can be guaranteed only when accurate training data is prepared or a very large volume of data is provided. In the following step of the research, the data presented in Table 7 will be analyzed and discussed. In addition, the errors on vowel will be identified for further analysis and discussion.

Acknowledgements

This work was supported by Hankuk University of Foreign Studies Research Fund of 2017.

References

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Finch, A., & Sumita, E. (2008). Phrase-based machine transliteration. In Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST) (pp. 13-18).

Finch, A., Dixon, P., & Sumita, E. (2012, July). Rescoring a phrase-based machine transliteration system with recurrent neural network language models. In Proceedings of the 4th Named Entity Workshop (pp. 47-51). Association for Computational Linguistics.

Finch, A., Liu, L., Wang, X., & Sumita, E. (2015, July). Neural network transduction models in transliteration generation. In Proceedings of NEWS 2015 The Fifth Named Entities Workshop (p. 61).

Hosseinzadeh, N. M., Kambuziya, A. K. Z., & Shariati, M. (2015). British and American phonetic varieties. Journal of Language Teaching and Research, 6(3), 647-655.

http://www.wildml.com/2016/08/rnns-in-tensorflow-a-practical-guide-and-undocumented-features/

Jurafsky, D. (2000). Speech & language processing. Pearson Education India.

Lee, K. J. and Choi, Y. S. 2017. Automatic Conversion of English Pronunciation Using Sequence-to-Sequence Model. *KIPS Transactions on Software and Data Engineering*, 6, 5, (2017), 267-278. DOI: 10.3745/KTSDE.2017.6.5.267

Sutskever, I., Vinyals, O., Le, Q.V. (2014). Sequence to sequence learning with neural networks. In NIPS-2014, pp. 3104—3112.

