KIPS Trans. Softw. and Data Eng.
Vol.8, No.4 pp.163~170 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2019.8.4.163

An LSTM Method for Natural Pronunciation Expression of Foreign Words in Sentences

Sungdon Kim† · Jaehee Jung††

ABSTRACT

Korea language has postpositions such as eul, reul, yi, ga, wa, and gwa, which are attached to nouns and add meaning to the sentence. When foreign notations or abbreviations are included in sentences, the appropriate postposition for the pronunciation of the foreign words may not be used. Sometimes, for natural expression of the sentence, two postpositions are used with one in parentheses as in "eul(reul)" so that both postpositions can be acceptable. This study finds examples of using unnatural postpositions when foreign words are included in Korean sentences and proposes a method for using natural postpositions by learning the final consonant pronunciation of nouns. The proposed method uses a recurrent neural network model to naturally express postpositions connected to foreign words. Furthermore, the proposed method is proven by learning and testing with the proposed method. It will be useful for composing perfect sentences for machine translation by using natural postpositions for English abbreviations or new foreign words included in Korean sentences in the future.

Keywords : Postposition, LSTM, Dropout, Overfitting, Final Consonant Pronunciation of Nouns

문장에 포함된 외국어의 자연스러운 발음 표현을 위한 LSTM 방법

김 성 돈† · 정 재 희††

요 약

한국어는 "을/를/이/가/와/과"와 같은 조사가 체언에 붙어 문장의 의미를 더해준다. 문장 중에 외국어 표기를 그대로 사용하는 경우나 외국어의 약자가 포함되어 있는 경우, 외국어의 발음에 따른 적절한 조사가 연결되지 않는 경우가 있다. 때로는 문장의 자연스러운 표현을 위하여 "을(를)"과 같이 괄호 형식으로 표현하여 조사를 두 개 다 수용 가능한 형태로 사용되어지기도 한다. 본 연구에서는 문장 내에 외국어가 포함되어 있는 경우, 조사가 부자연스럽게 연결되는 예를 찾고 체언의 종성 발음을 학습하여 자연스러운 조사 연결을 위한 방법을 알아보고자 한다. 제안하는 방법은 순환신경망 모델을 이용하여 외국어에 연결된 조사를 자연스럽게 표현하는 것이다. 제안된 모델로 학습 및 테스트하여 방법의 필요성을 입증함으로써, 향후 기계 번역에서 영문 약자나 새로운 외국어 삽입 시 자연스러운 조사 연결로 완전한 문장을 연결하는데 사용될 수 있을 것으로 기대한다.

키워드 : 조사, LSTM, 드롭아웃, 과적합, 종성

1. 서 론

조사는 다른 단어와 함께 쓰여 문장 내에서 다른 말과의 관계를 자연스럽게 연결해주는 품사를 의미한다. 한국어의 조사는 명사, 대명사와 같은 체언에 붙어 다른 말과의 관계를 잘 나타내준다. 예를 들어 "나 커피 좋아한다." 보다는 "나는 커피를 좋아한다." 와 같이 조사를 사용할 경우 그 의미를 쉽게 알 수 있다. 급속하게 변화는 사회에서 외래어의 사용 빈도도 높아지고, 외래어를 번역하지 않고 그대로 사용하는 것이 문맥상 더 잘 어울린다고 생각되기 때문에 음차어(Transliteration)를 그대로 사용하는 경우가 많아졌다[1]. 이에 따라 한국어로 문장을 구사할 때 외래어와 음차어에 조사를 연결하여 사용하는 빈도도 점차 늘고 있다.

하지만 음차어나 외래어를 그대로 한국어 문장에 사용하면서 발생하는 문제점은 적절하지 않은 조사가 사용되기도 한다. 이러한 오류를 줄이기 위하여 목적격 조사의 경우 "을

※ 이 논문은 한국연구재단의 지원을 받아 수행된 연구임(NRF-2016R1C1B1007929).
† 비 회 원 : 홍익대학교 정보컴퓨터학부 학사과정
†† 정 회 원 : 홍익대학교 교양학부 조교수
Manuscript Received : November 28, 2018
First Revision : February 8, 2019
Second Revision : March 7, 2019
Accepted : March 21, 2019
* Corresponding Author : Jaehee Jung(jhjung@hongik.ac.kr)

※ This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/ licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

(를)”과 같이 두 개의 조사를 동시에 사용되기도 한다. 예를 들어 영어에서 “Oil”은 “오일을(를)”으로 “Coffee”는 “커피을(를)”과 같이 사용한다.. “MAC”와 같은 고유명사도 한국어로 표기 시 “맥”과 같이 표기되기 때문에 “I like MAC”과 같은 문장은 “나는 맥을 좋아 합니다.” 라고 표기 되어 목적격 조사 “을”을 쓰는 것이 조금 더 자연스러운 표현이다. 하지만 구글(Google) [2], 카카오(Kakao)[3], 네이버(Naver)[4]의 번역에서는 맨 마지막 알파벳인 “C” (발음 “씨”)에 중점을 두어 “나는 MAC를 좋아한다.”로 “를” 목적격 조사가 사용되었다. 즉, 목적격 조사 이전에 위치하게 되는 명사 발음의 종성 여부에 따라 “을”과 “를”이 구분되어지지만 한국어 종성에 대한 발음 종성에 대한 발음 적용에 따른 조사 연구가 미흡함에 따라 “을”,“를”과 같은 정확한 조사 사용에 대한 학습이 이루어지지 않고 있다.

Table 1. The Examples of Automatic Translation by Google, Naver and Kakao Applications

Application	Example
Google	난 어려운 수준의 GMAT를 해결했습니다.
Naver	나는 높은 수준의 GMAT를 해결했다.
Kakao	나는 높은 난이도의 GMAT를 해결했다.

Table 1은 많이 사용되고 있는 기계 번역에 대한 세 가지 예로 “I solved high difficulty level GMAT”를 자동 번역한 문장이다. GMAT은 한국어로 “지맷”으로 발음되면서 종성에 받침소리가 있기 때문에 “GMAT을” 이라고 조사를 사용하는 것이 조금 더 자연스러운 표현이다. 하지만 “T“가 ”티”로 발음되어 종성 받침소리가 없기 때문에 “을”대신 “를”으로 자동 번역되었다.

최근에 활발히 진행되고 있는 연구 분야는 언어 음성 인식[5]이나 자연어 처리 자동 기계번역[6,7]에 적용되고 있는 순환신경망(Recurrent Neural Network – RNN) 모델이다. 순환신경망은 입력된 정보의 순서가 중요한 경우 사용되는 기계학습 신경망 중 하나이다. 본 연구에서는 문장 중에 외래어가 포함된 경우 단어의 문자의 순서를 학습하여 자연스러운 조사 연결을 위한 순환신경망 모델을 방법을 제안하고자 한다.

2. 방 법

2.1 데이터

데이터는 영어 참고사이트[9]에서 선택된 단어와 고유명사를 합하여 총 5,614개의 영어 단어를 이용하였다. 사용된 데이터에 대한 분포는 Table 2, 3과 같다.

Table 2는 단어의 맨 마지막 철자가 자음인지 모음인지를 나타내는 개수를 의미하며, 자음인지 모음인지에 따라 자연스럽게 연결될 수 있는 조사 선택에 영향을 미친다. Table 2에서 보는 바와 같이 자음으로 끝나는 단어가 절대적으로 많음을 볼 수 있다.

Table 2. The Number of Vowel and Consonant Word in Dataset

Last letter of word	No. data
Vowel	1420
Consonant	4194
Total	5614

Table 3은 데이터 5614개를 품사별로 구분하여 나타낸 표로 품사를 구분하기 위하여 NLTK: the Natural Language Toolkit [10]을 이용하여 품사를 구분하였다. 일반적으로 명사 뒤에 조사가 가장 많이 연결되어 사용되기 때문에 데이터의 비율 중 명사가 가장 많다.

Table 3. The Number of Parts of Speech in Dataset

Parts of speech	No. data
None	3825
Verb	642
Adjective	392
Adjective satellite	328
Adverb	313
etc.(conjunction, preposition etc)	114
Total	5614

2.2 조사 (“을”, “를”) 분류를 위한 전처리

학습 및 생성된 모델의 유효성 검증을 위하여 데이터를 구분 할 수 있는 구분이 필요하다. 영어 단어에 연결 될 수 있는 적절한 조사 선택을 위하여 한국어 발음으로 번역되어진 기존의 툴을 사용하였다. 영어 단어에 연결되는 조사를 자동으로 분류하기 위하여 사용되어진 방법은 깃허브 transliteration[8]으로 외국어 발음을 그대로 한국어 발음으로 음차 표기 할 수 있는 툴이다.

transliteration으로 음차표기 된 한국어 발음을 작성하고, 한글의 유니코드를 이용하여 해당 한국어 발음의 마지막 글자에 받침이 있는지 없는지 확인하였다. 자동으로 종성의 자음 받침의 유무에 따라 두 개의 클래스로 구분 하였으며,“을”이 사용되는 경우는 1로 정의하고 “를”이 사용되는 경우는 0으로 정의 하였다. 예를 들어 transliteration을 이용한 “above”의 한국어 발음은 “아보브”이다. 이때 “브”의 유니코드(Unicode)를 확인하면 종성이 없다. 따라서 above의 마지막 발음은 자음 받침이 없다고 판단 후, 자동으로 유니코드를 추출하여 마지막 발음에 자음이 받침이 있는지 없는지에 따라 데이터의 클래스를 구분하였다.

하지만 Table 2에 변환된 예에서 “제안된 한국어 발음” 필드에서 보는 바와 같이 Transliteration에 의해 자동으로 생성된 발음 중 몇 단어는 현실과는 다른 발음으로 발음되기 때문에 종성이 있어야 할 한국어 발음에 종성이 없는 경우도 발생한다. 예를 보면 “big”과 같은 단어가 “비그”로 발음을 제안하였다. 하지만 “비그” 대신 “빅”이라는 발음으로 주로 쓰이며

단어 종성의 받침이 존재 유무에 따라 다른 목적격 조사가 사용될 수 있다. 잘 못 음차 표기된 경우에 한하여 오류 수정이 필요하다. Transliteration에서 음차표기 오류로 생각되어지는 단어는 사용된 5614개의 데이터 중 115개 이고, 오류가 있는 음차 표기에 대해서는 수동으로 수정하여 조사 사용의 적절한 클래스 구분으로 사용 되었다. 적절한 조사 클래스의 전처리를 위하여 5614개의 단어 데이터를 Transliteration에서 오류가 없는 경우는 Table 5의 두 번째 열을 사용하였고, 오류가 있는 경우 맨 마지막 열에서 보는 바와 같이 생성 하였다. 종성에 받침이 있는지 없는지를 자동으로 구분하여 각 단어 별로 “0” 과 “1”을 다른 클래스로 구분하여 두 개의 클래스를 생성하였다.

Table 4. The Examples of Wrong Transliteration in Korean

English Word	Suggested Korean Transliteration	Error (Correct Pronounce)
abroad	아브로드	
actual	악투	액투얼
alone	알론	
big	비그	빅
but	부트	벗
claim	클레임	
conference	콘퍼런스	
deep	디프	딥
drop	드롭	
mosaic	모자	모자익
poet	포트	포엣

단어 데이터는 “을”과 “를”로 분류된 클래스는 총 2종류 이고, 각각의 개수는 Table 6과 같다. 전체 데이터의 비율에서 각 클래스의 비율을 보았을 때 “을”의 비율이 32%이고 “를”의 비율이 68%로 “을”의 데이터가 상대적으로 “를”의 비율보다 낮게 나타나고 있다.

2.3 학습 및 테스를 위한 단어의 전처리

사용될 단어는 Fig. 1과 같이 데이터를 단어의 문자열 순서를 역순으로 전 처리하였다. 적절한 조사의 선택은 발음에서의 단어의 종성에 따라 분류가 달라지는 점에 착안하여 단어를 배치할 때 Fig. 1과 같이 최대 길이를 설정하고 영어 단어 문자열을 역순으로 입력하였다. 이때 역순의 최대 5개의 문자열만 학습을 하도록 하고 5개의 문자열이 넘는 경우에 대해서는 사용하지 않았다. 이유는 단어의 앞에서 발음되고 있는 모음 및 자음은 조사의 형태, 즉 “을” 또는 “를”의 합성에 영향을 미치지 않기 때문이다. Fig. 1에서는 채워진 파란색 5개의 영역이 학습을 하도록 시킨 것을 의미하고, 빨간색 점선으로 된 부분은 전체 길이가 5보다 길기 때문에 고려하지 않았다. 전제 단어의 문자열이 단어의 최대 문자인 5개가 넘지 않는 경우에 대해서는 마지막 인덱스에 가까운 부분에 NULL 값을 임의로 입력하였다. 각각의 문자는 원-핫-인코딩(one-hot-encoding)으로 저장되어 순환신경망의 입력 값으로 사용된다. Fig. 1의 맨 아래부분은 'R' 을 표현하고 있고, 역순으로 저장된 5개의 문자 역시 각각의 원-핫-인코딩으로 표현하여 순환신경망의 입력 값으로 사용하였다. Fig. 1의 노란색으로 표현된 부분은 2.2절에서 설명한 바와 같이 조사의 클래스를 의미하며 “을”은 1로 표현하였고 “를”은 0으로 표현하였다.

Table 5. Data Classification Depending on the Korean Pronunciation. “1” at postposition class stands for “eul - 을” and “0” means “reul-를”

English word	Korean Transliteration	Postposition Class
able	에이블	1
air	에어	0
ball	볼	1
bench	벤치	0
book	북	1
calorie	칼로리	0
carpet	카펫	1
can	캔	1
concert	콘서트	0
gas	가스	0
log	로그	0

Table 6. The Number and Distribution of Data for Each Class

Postposition Class	No. data
을 eul (1)	1774 (Around 32 %)
를 reul (0)	3840 (Around 68 %)
Total	5614

Table 7은 단어의 길이에 따른 개수를 의미한다. Fig. 1에서 도식한 바와 같이 최대 5개까지의 문자만 유의미한 학습 및 테스트 데이터로 사용하였기 때문에 단어의 길이가 3, 4, 5, 6 이상일 경우에 대하여 구분하여 해당 개수를 표로 나타내었다.

Table 7. The Number of Word for Each Length in Dataset

Length	No. data
3	241
4	703
5	802
Over 6	3868
Total	5614

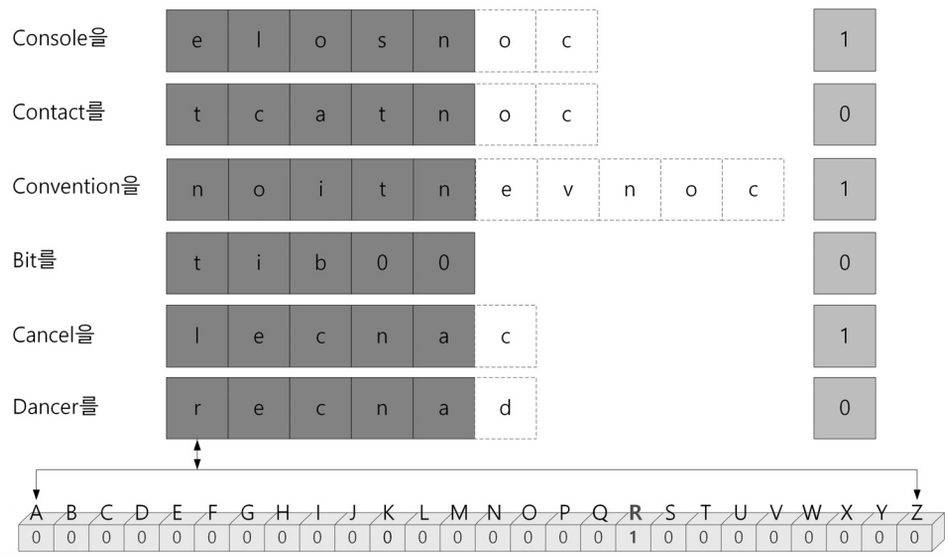

Fig. 1. Data Composition by One-hot Encoding with Last Five Characters of the Word

3. 실험 및 결과

총 5614개의 데이터 중 5%를 테스트로 사용하여 약 280개를 테스트로 사용하고 5534개를 학습 데이터로 사용하였다. 교차 검증(cross validation)을 위하여 10번을 임의로 나누어 테스트의 평균 정확도를 평균으로 계산하였다.

3.1 모델구성

1) 실험환경

실험을 위해 텐서플로우(Tensorflow)[11] 프레임 워크를 사용하였다. 텐서플로우는 Theano[12], Torch[13], Caffe[14]와 더불어 가장 많이 사용되고 있는 딥러닝 프레임워크(Deep Learning Framework)로 모델을 쉽게 생성할 수 있고, 텐서보드(Tensorboard)라는 인터페이스를 제공하여 모델을 시각화하고 쉽게 최적화 할 수 있다는 장점을 지닌다. 실험 환경은 우분투 16.04 환경에서 1.6.0 버전으로 GeForce GTX 1080 Ti 스펙의 GPU가 설치된 환경에서 실험하였다.

2) Long Short Term Memory (LSTM)

문자 음성과 같은 순차적인 데이터가 데이터의 클래스 결정에 영향을 미치는데 사용되는 방법은 순환신경망 방법이다. 이전 단계에서 숨겨진 노드(Hidden Node)가 다음 단계의 입력 값으로 사용되는 것이 다른 신경망과의 차이이다. 하지만 여러 단계를 숨겨진 노드를 계산하면 결정적으로 영향을 미치는 정보들이 점차 그 영향력이 줄어듦에 따라 학습능력이 저하되는 단점이 발생한다. 이를 극복하기 위한 방법이 LSTM 방법으로 순환신경망에 숨겨진 단계에 셀 구조를 추가한 방법이다. 제안된 방법은 Stacked Long Short Term Memory (Stacked LSTM)을 사용하였으며 스택의 맨 아래는 드롭아웃(Dropout) 기법을 사용하였다. 드롭아웃은 과적합(Overfitting)을 막기 위한 기법 중 하나이다. 과적합은 학습 데이터 셋에 지나치게 학습되어 테스트 데이터셋으로 테스트하였을 때 오히려 그 학습 데이터의 정확도가 떨어지는 상태를 말한다.

드롭아웃의 성능 비교를 위하여 세 개의 층으로 쌓여진 LSTM 방법을 서로 다른 세 가지 방법으로 비교 분석하였다. Fig. 2 (A) 그림은 드롭아웃을 한 계층(Layer)도 실행하지 않은 방법을 의미하고, Fig. 2 (B) 그림은 첫 번째 계층만 드롭아웃을 실행한 경우, Fig. 2 (C) 그림은 모든 계층에 드롭아웃을 실행한 경우에 대한 비교이다. Fig. 2에서 빨간색 선은 학습 데이터를 이용한 정확도를 파란색 점선은 테스트 데이터를 이용한 정확도를 의미한다. 두 그래프 간의 격차가 클수록 학습 데이터에 모델이 과적합, 즉 학습 데이터의 특징들에 지나치게 학습되어 모델이 생성되었음을 의미한다. Fig. 2 (A) 그림에서 보는 바와 같이 드롭아웃을 시행하지 않았을 경우 빨간색으로 표시된 학습 데이터 정확도와 파란색으로 표시된 테스트 데이터 정확도가 다른 Fig. 2 (B), (C) 그래프에 비해서 격차가 나는 것을 알 수 있다. 제안된 모델에서는 드롭아웃 방법을 사용하는 것이 드롭아웃을 사용하지 않는 방법보다 학습데이터에 적게 과적합되어 학습 모델에 유의미하게 영향을 미치는 것을 알 수 있다. 즉, 드롭아웃을 설정하는 것이 학습 데이터에 과적합되지 않는 모델을 생성할 수 있다.

Fig. 3는 드롭아웃 방법이 성능 향상에 영향을 미치는 결과를 기반으로 하여 계층을 3개-Fig. 3 (A), 4개-Fig. 3 (B), 5개-Fig. 3 (C)로 각각 설정하고 모든 계층에 드롭아웃을 기법을 적용하였을 때에 대한 성능 비교를 하였다. Fig. 3에서 보는 바와 같이 계층을 추가함에 따라 성능이 둘의 차이가 미세하게 줄어들지만, 테스트 데이터와 학습 데이터의 정확도가 크게 차이가 없어 드롭아웃의 계층 수가 과적합에 큰 영향을 주지 않음을 보였다.

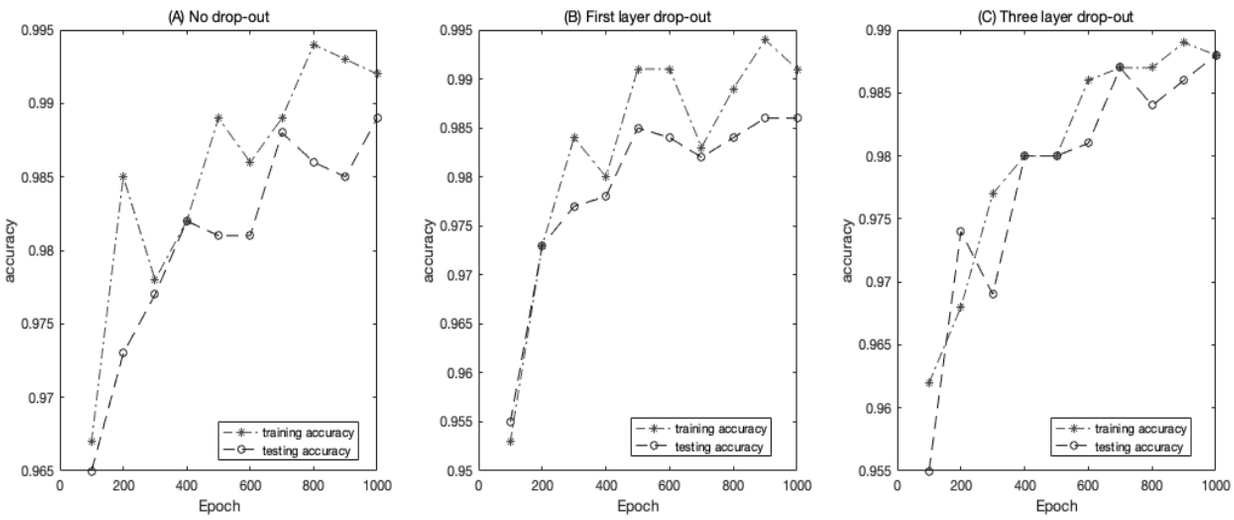

Fig. 2. The Accuracy with Training and Testing Data Set.
(A) Drop-out technique is not applied on any three stacked layers, (B) Drop-out technique is applied on the only bottom layer among three stack layer, (C) Drop-out technique is applied all three stack layers

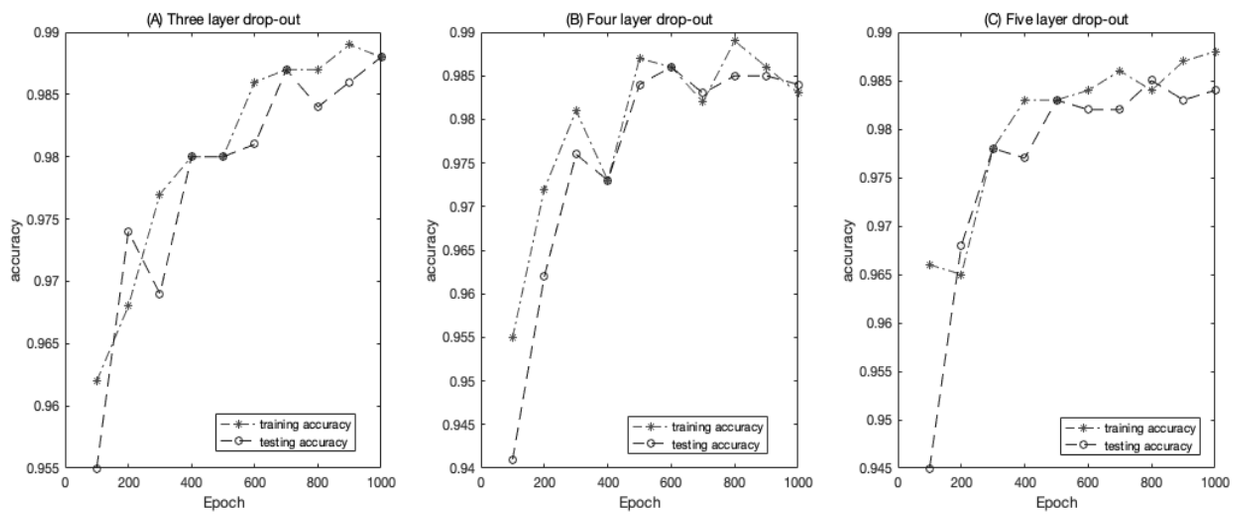

Fig. 3. The Accuracy with Training and Testing Data Set.
(A) Drop-out technique is applied all three stack layers, (B) Drop-out technique is applied all four-stacked-layers, (C) Drop-out technique is applied all five stack layers

따라서 제안하는 방법은 Fig. 4와 같이 설계되었으며 5개의 Stacked 계층 중 맨 아래 계층만 128개의 숨겨진 셀(Hidden Cell)을 생성할 때 드롭아웃의 비율을 0.5로 설정하였고, 그림에서 짙은 초록색으로 표현하였다. 드롭아웃을 설정하지 않은 다른 계층은 128개의 숨겨진 셀을 그대로 사용하였으며 옅은 초록색을 구분하였다. 파란색 원은 입력된 원은 Fig. 1과 알파벳이 원-핫-인코딩으로 변형된 데이터를 의미한다. 128개의 숨겨진 계층(Hidden layer)을 구성할 때 5개의 쌓여진(Stacked) 계층을 구성하였으며, 드롭아웃한 셀은 진한 초록색으로 그렇지 않은 계층(Layer)은 옅은 초록색으로 표시하였다. 각 계층에서의 저장된 값을 분홍색으로 표현하였으며 이 값을 SoftMax 값으로 하여 1(을) 또는 0(를) 클래스로 예측할 수 있다.

일반적인 순환신경망 모델은 한 단계를 실행하면 어떠한 결과를 예측할 수 있고 그 결과는 $y = wx + b$와 같이 표현될 수 있다. Fig. 5는 텐서보드를 이용한 모델의 형태를 보여주고 있다. 일반적으로 w는 가중치를 의미하고 b는 오차를 의미한다. 가중치와 기울기를 임의의 랜덤 값으로 설정하여 단계가 증가할 때마다 w와 b를 비용이 적게 발생하는 방향으로 수정하여 모델을 생성하게 된다. 따라서 Fig. 5에서의 텐서 보드에서 보듯이 왼쪽 중간에 있는 variable이 b를 Variable_1가 w를 의미한다.

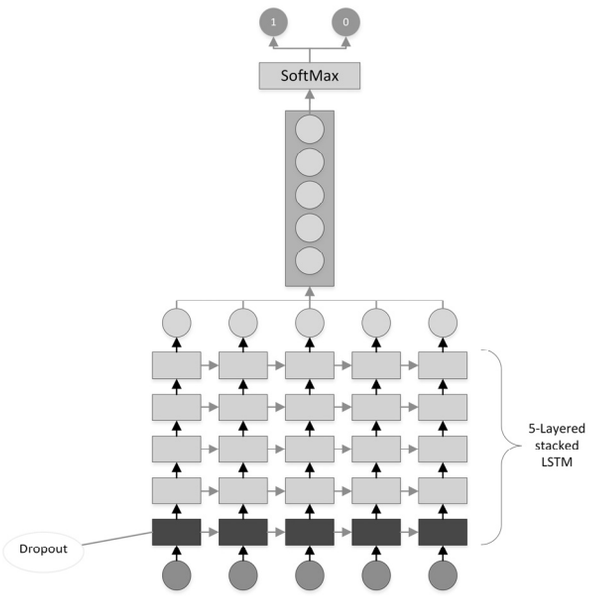

Fig. 4. The Framework of Suggested Model

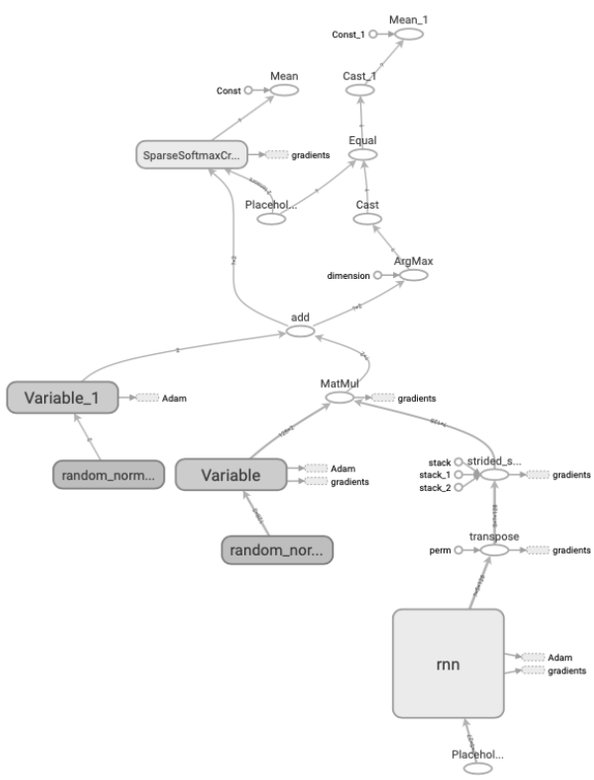

Fig. 5. The Screen-shot of Tensor Board for the Suggested Model

3.2 결과분석

Fig. 6은 각 Epoch 마다의 평균 정확도를 보여주고 있다. 빨간색 점은 평균 정확도를 의미하고 위아래의 파란색 범위의 오차는 10번을 임의로 데이터 셋을 나누어 교차 검증하였기 때문에, 정확도의 최댓값과 최솟값을 의미한다. Epoch가 진행이 되면서 점점 정확도가 높아지는 것을 알 수 있고, 특히 1000번에 가까울수록 정확도가 거의 1에 수렴하는 것을 볼 수 있다. 약 500번째의 Epoch에서 평균 정확도가 확연히 떨어지는 것은, 학습 데이터의 과적합으로 인한 테스트 데이터의 정확도가 떨어지는 것으로 판단된다.

결과 분석에 Receiver Operating Characteristics (ROC)도 사용하였다. Table 7에서 보는 것과 같이 두 개의 클래스의 사이즈가 완전한 균형을 이루고 있지 않다. “을”을 1로 그룹화 하였고 “를”을 0으로 그룹화 하였는데 약 1:2 의 비율로 “를”의 클래스가 더 많은 것을 보여주고 있다. 하지만 두 클래스의 개수가 균형이 맞지 않을 경우나 한 클래스 데이터가 지나치게 많이 학습될 경우 정확도 지표는 높아지지만 모델의 정확성을 판단하기에 어렵기 때문에 ROC를 이용하여 비교하였다 (Fig. 7). ROC curve는 MATLAB의 파일[15]을 이용하였으며, x축은 1 – Specificity = FP / FP + TN를 의미하고 y 축은 Sensitivity = TP / TP + FN를 의미한다. Fig. 7은 제안된 방법 이외에 Support Vector Machine (SVM) 과 로지스틱 회귀분석(Logistic Regression)을 비교하였으며, 제안된 모델이 더 높은 ROC 커브를 보여준다.

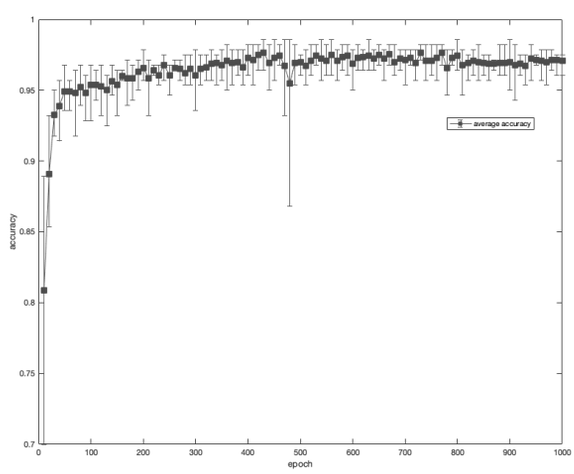

Fig. 6. The Min, Max, and Average Accuracy for Each Epoch. The Red Point Stands for Average of Accuracy and Blue Range is Min and Max of Accuracy

이때 사용되어진 Sensitivity와 Specificity는 Table 8의 confusion matrix를 근거로 한다. confusion matrix는 두 개의 클래스로 나누어질 때의 분류 결과표를 의미한다. 표를 기반으로 정확도는 TP + TN / (TP + FN + FP + TN) 로 나타낼 수 있고 Fig. 6으로 도식화하였다. Table 8에서 괄호 안에 있는 숫자는 10번의 교차 검증 중에 평균으로 구해진 TP, FP, TN, FN를 의미하고 Fig. 8은 각 10번의 교차 검증으로 얻어진 TP, FP, TN, FN의 평균을 매 epoch마다 그림을 의미한다. 매 회당 TN 가 TP 보다 훨씬 많은 비중으로 차지하는 것으로 나타났고, Fig. 6과의 상관성으로 약 500번째의 epoch에서 TN와 FP의 값이 상대적으로 더 올라가거나 내려가는 것을 확인할 수 있다. Fig. 6, Fig. 7, Fig. 8을 기반으로 영어단어의 문자 순서를 학습하여 의미 있는 조사 선택이 가능함을 알 수 있다.

Table 8. Confusion Matrix

	Predicted : YES	Predicted : NO
Actual YES	True Positive(TP) (94.3)	False Negative(FN) (6.67)
Actual NO	False Positive(FP) (3.96)	True Negative(TN) (174.04)

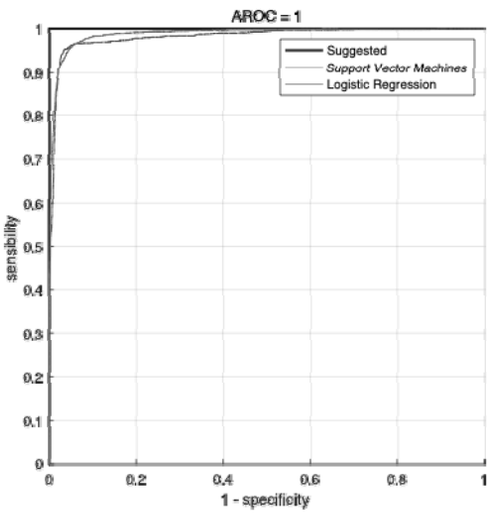

Fig. 7. The ROC Curve

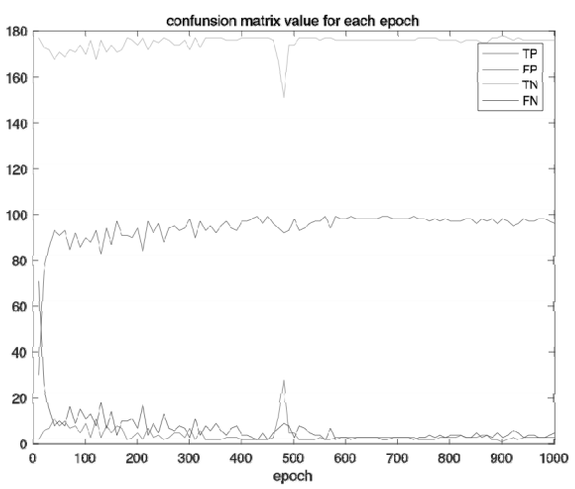

Fig. 8. Average Confusion Matrix Values of 10-fold Cross Validation for Each Epoch

4. 결 론

한국어는 조사 포함 여부에 따라 의미가 다르게 해석될 수 있기 때문에 적절한 조사의 사용이 매우 중요하다. 최근에는 외래어를 번역하지 않고 음차어를 그대로 사용하는 경우나 고유한 의미의 외래어를 그대로 사용하는 빈도가 높아짐에 따라 의미를 적절하게 연결해주는 올바른 조사의 선택이 필요하다.

본 논문에서는 순환신경망의 방법 중 하나인 LSTM을 이용하여 외래어를 음차표기 후 뒤에 연결되는 조사를 한국어 발음에 따라 학습하는 방법을 제안하였다. 데이터는 외래어를 한국어 발음으로 변환 후 종성의 유니코드 판별로 클래스를 구분 하였으며, 외래어나 번역이 되지 않는 고유명사도 포함하였다. 발음에 맞는 적절한 조사를 이용하여 LSTM 방법으로 학습 및 테스트하고 교차 검증으로 과적합 유무를 판단하였다. 학습 방법은 스택 계층을 쌓고 첫 번째 계층만 드롭아웃을 사용한 텐서플로우의 LSTM 이용하였으며, 발음에 적합한 조사를 자동으로 연결하기 위하여 ROC성능 비교를 이용하여 다른 패턴인식 방법과 비교하여 우수함을 입증하였다.

제안된 모델은 기존의 Transliteration과 같은 음차 표기 방법을 기반으로 종성 유무와 한국어의 조사를 선택한 것과는 달리 영어와 같은 외래어를 문자의 순서로 학습하여 적절한 조사 선택에 관한 모델을 제시하였다. 클래스를 구분하기 위하여 github의 Transliteration을 사용하여 자동으로 “을”,“를”을 구분하는데 사용하였으나, 학습 모델은 문자열 자체로 학습하여 음차표기의 정확도와 별개로 단어 문자 배열로 적절한 조사를 제시할 수 있다. 이는 향후 외래어나 음차표기를 그대로 한국어 문장에 삽입 시 자연스러운 조사 선택에 도움이 될 수 있을 것으로 기대 한다.

하지만 이 실험은 적절한 조사 클래스 구분을 위하여 영어를 기반으로 한 한국어 음차 표기 방법인 github Transliteration을 기준으로 클래스를 구분하였다. 음차표기의 오류가 진행된 부분은 수동으로 정정되었기 때문에 영어를 제외한 다른 외국어의 자동으로 정확한 조사의 구분에 한계를 갖는다. 향후 추가적으로 진행되어야 할 연구는 외래어 발음에 따른 적절한 조사 선택을 위하여 음차 표기의 정확성을 높이는 연구이다.

References

[1] Songyi Lee, “A Study on Perception of English-Transliteration Words in Newspaper Articles,” *Studies in Linguistics*, Vol.46, No.1, pp.313-333, 2018.

[2] Google Translation, [Online]. Available: https://translate.google.com

[3] Naver Translation, [Online]. Available: https://papogo. naver.com

[4] kakao Translation, [Online]. Available: https://translate.kakao.com

[5] Lee, Donghyun, Lim, Minkyu, Park, Hosung, and Kim, Ji-Hwan, “LSTM RNN-based Korean Speech Recognition System Using CTC,” *Journal of Digital Contents Society*, Vol.18, No.1, pp.93-99, 2017.

[6] Goldberg, Yoav, “A Primer on Neural Network Models for Natural Language Processing,” *Journal of Artificial Intelligence Research*, Vol.57, No.1, pp.345-420, 2016.

[7] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al., “Google’s neural machine translation system: Bridging the gap between human and machine translation,” arXiv preprint arXiv:1609.08144, 2016.

[8] English-Korean Transliteration [Internet], https://github.com/muik/transliteration
[9] WordNet: A Lexical Database for English [Internet], https://wordnet.princeton.edu/
[10] Edward Loper and Steven Bird, "NLTK: the Natural Language Toolkit," *ETMTNLP '02 Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, Vol. 1, pp.63-70.
[11] TensorFlow Release [Internet], https://www.tensorflow.org/, Retrieved 14 November 2018.
[12] Theano Release [Internet], http://www.deeplearning.net/software/theano/, Retrieved 17 September 2018.
[13] R. Collobert, S. Bengio, and J. Marithoz, "Torch: a modular machine learning software library," Technical Report IDIAP-RR02-46, IDIAP, 2002.
[14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama and Darrell, Trevor, "Caffe: Convolutional Architecture for Fast Feature Embedding," 2014.
[15] Víctor Martínez-Cagigal, ROC Curve [Internet], (https://www.mathworks.com/matlabcentral/fileexchange/52442-roc-curve), MATLAB Central File Exchange. Retrieved February 7, 2019.

김 성 돈

https://orcid.org/0000-0002-7993-0926
e-mail : deltacluse@naver.com
2017년~현 재 홍익대학교 정보컴퓨터학부 학사과정
관심분야 : 머신러닝, 강화학습, 데이터분석, 자연어처리

정 재 희

https://orcid.org/0000-0002-0932-3039
e-mail : jhjung@hongik.ac.kr
2000년 동덕여자대학교 전산통계학과(학사)
2002년 고려대학교 컴퓨터학과(석사)
2008년 Texas A&M University Computer Science(Ph.D)
2009년~2014년 삼성전자 책임연구원
2015년~현 재 홍익대학교 교양학부 조교수
관심분야 : Bioinformatics, Pattern Recognition, Machine Learning