KIPS Trans. Softw. and Data Eng.
Vol.8, No.11 pp.441~448 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2019.8.11.441

Automatic Word Spacing of the Korean Sentences by Using End-to-End Deep Neural Network

Hyun Young Lee† · Seung Shik Kang††

ABSTRACT

Previous researches on automatic spacing of Korean sentences has been researched to correct spacing errors by using n-gram based statistical techniques or morpheme analyzer to insert blanks in the word boundary. In this paper, we propose an end-to-end automatic word spacing by using deep neural network. Automatic word spacing problem could be defined as a tag classification problem in unit of syllable other than word. For contextual representation between syllables, Bi-LSTM encodes the dependency relationship between syllables into a fixed-length vector of continuous vector space using forward and backward LSTM cell. In order to conduct automatic word spacing of Korean sentences, after a fixed-length contextual vector by Bi-LSTM is classified into auto-spacing tag(B or I), the blank is inserted in the front of B tag. For tag classification method, we compose three types of classification neural networks. One is feedforward neural network, another is neural network language model and the other is linear-chain CRF. To compare our models, we measure the performance of automatic word spacing depending on the three of classification networks. linear-chain CRF of them used as classification neural network shows better performance than other models. We used KCC150 corpus as a training and testing data.

Keywords : Syllable Embedding, Bi-LSTM, Feedforward Neural Network, Neural Network Language Model, Linear-Chain CRF

종단 간 심층 신경망을 이용한 한국어 문장 자동 띄어쓰기

이 현 영† · 강 승 식††

요 약

기존의 자동 띄어쓰기 연구는 n-gram 기반의 통계적인 기법을 이용하거나 형태소 분석기를 이용하여 어절 경계면에 공백을 삽입하는 방법으로 띄어쓰기 오류를 수정한다. 본 논문에서는 심층 신경망을 이용한 종단 간(end-to-end) 한국어 문장 자동 띄어쓰기 시스템을 제안한다. 자동 띄어쓰기 문제를 어절 단위가 아닌 음절 단위 태그 분류 문제로 정의하고 음절 unigram 임베딩과 양방향 LSTM Encoder로 문장 음절간의 양방향 의존 관계 정보를 고정된 길이의 문맥 자질 벡터로 연속적인 벡터 공간에 표현한다. 그리고 새로이 표현한 문맥 자질 벡터를 자동 띄어쓰기 태그(B 또는 I)로 분류한 후 B 태그 앞에 공백을 삽입하는 방법으로 한국어 문장의 자동 띄어쓰기를 수행하였다. 자동 띄어쓰기 태그 분류를 위해 전방향 신경망, 신경망 언어 모델, 그리고 선형 체인 CRF의 세 가지 방법의 분류 망에 따라 세 가지 심층 신경망 모델을 구성하고 종단 간 한국어 자동 띄어쓰기 시스템의 성능을 비교하였다. 세 가지 심층 신경망 모델에서 분류 망으로 선형체인 CRF를 이용한 심층 신경망 모델이 더 우수함을 보였다. 학습 및 테스트 말뭉치로는 최근에 구축된 대용량 한국어 원시 말뭉치로 KCC150을 사용하였다.

키워드 : 음절 임베딩, 양방향 LSTM, 전방향 신경망, 신경망 언어 모델, 선형 체인 CRF

1. 서 론

스마트폰의 대중화로 메신저, SNS, 블로그 등에 텍스트 입력이 간편해지면서 띄어쓰기 오류가 발생하는 경우가 증가하고 있다. 띄어쓰기 오류는 한국어 문장을 해석하는데 문법적, 의미적, 모호성을 발생시킨다. 띄어쓰기가 잘 적용된 이상적인 한국어 문장 기반으로 연구해온 기존의 형태소 분석, 구문 분석, 개체명(NER) 인식과 같은 자연어 처리 시스템은 띄어쓰기 오류를 내포하고 있는 한국어 문장 처리에 어려움이 있다[1]. 띄어쓰기 오류를 교정하는 것은 자연어 처리 시스템의 성능 향상에 영향을 미치는 요소이다. 따라서 한국어 문장에 대한 자연어 처리 시스템의 전처리 과정으로 자동 띄어쓰기

※ 이 논문은 2017년 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017M3C4A7068186).
† 준 회 원 : 국민대학교 컴퓨터공학과 박사과정
†† 종신회원 : 국민대학교 소프트웨어학부 교수
Manuscript Received : February 12, 2019
First Revision : April 12, 2019
Second Revision : August 2, 2019
Accepted : September 20, 2019
* Corresponding Author : Seung Shik Kang(sskang@kookmin.ac.kr)

를 적용하고 띄어쓰기 오류를 교정한 올바른 어절 생성이 필요하다.

한국어 자동 띄어쓰기는 음절과 음절 사이에 공백을 삽입하는 문제로 부분적 띄어쓰기 오류를 내포하는 문장과 띄어쓰기가 전혀 되어있지 않은 문장과 같이 두 가지 형태로 나누어진다[1,2]. 부분적 띄어쓰기 오류는 복합명사 분해, 철자 교정 등의 2-3 어절에 걸친 띄어쓰기 교정으로 이는 공백을 제거하여 전혀 띄어쓰기가 적용되지 않은 형태로 변환한다. 그리고 음절 사이에 공백을 삽입하여 공백이 존재하지 않는 문장의 띄어쓰기 오류 수정과 같은 방법론을 적용하여 띄어쓰기 오류를 수정할 수 있으므로 본 논문에서는 띄어쓰기가 전혀 되어있지 않은 문장을 대상으로 Fig. 1과 같이 심층 신경망을 이용한 한국어 문장의 자동 띄어쓰기 방법을 제안한다.

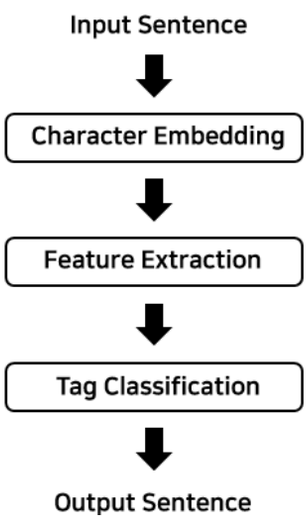

Fig. 1. Korean Automatic Word Spacing by Deep Neural Network

2. 관련 연구

기존의 한국어 자동 띄어쓰기에 관한 연구는 규칙 기반과 확률 및 통계적 정보를 이용한 방식으로 나누어진다. [1]은 말뭉치로부터 어절 빈도 사전과 음절 빈도 사전 기반의 어절 정보와 형태소 분석기를 이용하여 자동 띄어쓰기를 한다. [2]는 1음절 사전, 2음절 사전, 말뭉치 기반으로 휴리스틱 정보와 형태소 분석기를 이용하여 띄어쓰기 오류 교정 후 어절 간의 재결합으로 3단계를 거쳐 띄어쓰기 오류를 교정한다. [3]은 조사/어미의 음절 특성과 조사/어미 사전을 이용하여 어절 블록 단위로 문장을 분할하고, 분할된 어절블록 내에서 양방향 최장일치법과 형태소 분석을 적용하여 띄어쓰기를 교정한다. 이 연구들은 빈도 기반의 사전을 구축한 후 사전 정보를 이용함과 더불어 추가적인 형태소 분석을 수행하여 어절 경계면에 공백을 삽입하는 형태로 띄어쓰기 오류를 수정한다. 사전과 형태소 분석 활용하는 기존의 연구에서 명사 사전(어절 단위)을 활용하는 경우에는 고유명사, 신조어, 외래어 같은 미등록어 처리에 어려움이 있다[1-3].

통계적 정보 등을 이용하는 기계학습 방법론으로는 structural SVM, 심층 신경망 등이 있다. [4]는 structural SVM의 모델이 분석한 결과와 사용자가 입력한 띄어쓰기 정보를 이용하여 띄어쓰기를 교정한다. [5]는 심층 학습 모델인 GRU-CRF와 명사 사전을 이용하여 띄어쓰기를 교정한다. [6]은 LSTM 기반 인코더와 디코더 형태의 Sequence-to-Sequence 모델을 이용하여 인코더에 의해 표현된 벡터를 가지고 디코더에서는 입력 문장의 각 음절에서 띄어야 하는지 아닌지를 예측하는 방식으로 띄어쓰기를 교정한다.

[7]은 입력 문장의 ngram 임베딩을 위해 1차원 컨볼루션(convolution)과 GRU를 사용하여 띄어쓰기를 교정한다. 이러한 기계학습 방법론은 띄어쓰기를 태그 열 부착 문제로 보고 띄어쓰기 오류를 교정한다[4, 5]. 그리하여 자동 띄어쓰기 문제를 Fig. 2와 같이 한국어 문장의 각 음절에 띄어쓰기 태그를 부착하는 문제로 정의하고 띄어쓰기 오류를 수정한다. 띄어쓰기 태그로는 두 개의 B(beginning) 태그, I(inside) 태그를 사용한다. B 태그는 어절의 첫음절 시작을 의미하고, I 태그는 어절에서 첫음절을 제외한 나머지 부분을 나타낸다. 문장의 띄어쓰기 오류 교정을 위해 공백이 존재하지 않는 문장을 대상으로 음절 태그 분류 후, B 태그 앞에 공백을 삽입하는 형태로 자동 띄어쓰기를 하였다.

심층 신경망 모델의 경우에는 입력 자질인 단어를 연속적인 벡터 공간에 표현한다[8-11]. 한국어의 단어는 음절의 조합이고 한국어의 초성, 중성, 종성의 경우의 수를 모두 고려해도 음절의 수는 11,172개이다. 이러한 특성 때문에 음절 단위 임베딩이 사전에 등록되지 않은 단어를 처리하는데 단어 단위 임베딩보다 유연하다. 그리하여 심층 신경망의 입력 자질로 음절 unigram을 연속적인 벡터공간에 표현하고, 음절 unigram 벡터들 간의 의존성을 부여한 새로운 자질 벡터를 연속적인 벡터 공간에 표현하였다. 그리고 각 문장의 음절을 BI 태그로 분류하여 자동 띄어쓰기를 수행하는 종단 간 심층 신경망 시스템을 제안한다.

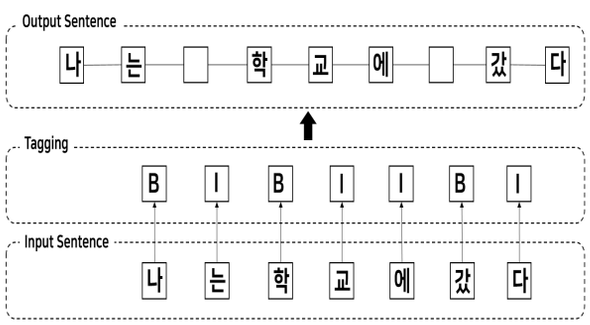

Fig. 2. Automatic Word Spacing of Korean Sentence with BI Tagging

3. 종단 간 심층 신경망을 이용한 자동 띄어쓰기

3.1 양방향 LSTM Encoder를 통한 자질 벡터 생성

심층 신경망 모델은 기존의 특징 추출과 패턴 분류의 두 가지 방식으로 나누어 해결하던 방식을 하나로 통합하여 문제를 해결하는 종단 간 시스템이다. 자연어 처리에서도 심층 신경망 모델은 문장을 구성하는 단어 등을 하나의 자질 정보로 보고 연속적인 벡터 공간에 단어를 표현한다. 벡터 공간 모델을 사용하는 심층 신경망 모델에서 텍스트를 연속적인 벡터 공간에 벡터로 표현하는 방식은 시스템 향상에 영향을 미치는 중요한 요소이다. 하지만 하나의 단어 벡터만으로는 문장에 속한 단어들 사이의 문맥 정보를 벡터 공간에 표현하는데 어려움이 있다. 그리하여 심층 신경망을 이용하는 자연어 처리에서는 문장에 속하는 단어를 순차적인 자질로 보고, LSTM을 이용하여 상호간의 의존성을 고려한 문맥 정보를 연속적인 벡터 공간에 표현한다. 이러한 LSTM는 문장에 속하는 단어 간의 의존성을 고려하는 방향에 따라 단방향 LSTM과 양방향 LSTM 형태의 두 가지 형태로 분류할 수 있다.

단방향 LSTM의 경우에는 현재 자질과 과거 자질의 문맥 의존성을 고려하여 문맥 정보를 연속적인 벡터 공간에 표현한다[12]. 현재 자질에 과거의 자질과의 의존성을 고려하는 단방향 LSTM은 과거의 문맥 정보만을 벡터로 표현하지만 양방향 LSTM은 과거 자질 뿐만 아니라 미래 자질도 현재 자질과 의존성을 고려하여 현재, 과거, 그리고 미래 자질들을 고려한 문맥 정보를 벡터로 표현한다[13, 14]. 예를 들어, 단방향 LSTM에서는 Fig. 3의 입력문장에서 현재 입력음절인 “학”이라는 음절은 “나”, “는”이라는 과거 문맥 정보와의 의존성만을 고려한다. 하지만 Fig. 3과 같이 양방향 LSTM의 현재 입력 음절인 “학”이라는 음절은 정방향 LSTM 셀로부터 “나”, “는”이라는 과거 음절 문맥 정보와 역방향 LSTM으로부터 “교”, “에”, “갔”, “다”라는 미래 문맥 정보를 얻고 현재 입력 음절 “학”이라는 음절 정보를 과거와 미래의 문맥 정보와 의존성을 고려하여 하나의 새로운 문맥 자질 벡터로 표현한다. 본 연구에서는 양방향 LSTM으로 음절간의 의존성을 표현한 새로운 자질 벡터를 Fig. 3과 같이 생성하고, BI 태그 분류를 위해 전방향 신경망(feedforward neural network), 신경망 언어 모델(neural network language model), 선형 체인(linear-chain) CRF의 분류 망에 따른 세 가지 모델을 구성하고 한국어 문장의 각 음절의 띄어쓰기 유형을 BI 태그로 분류한다.

3.2 전방향 신경망을 이용한 자동 띄어쓰기

전방향 신경망은 층마다의 노드들 사이에 사이클(cycle)이 존재하지 않고 입력 값에서 출력 값으로 한 방향으로 입력 자질 정보를 전달한다. 또한 전방향 신경망은 고정된 길이의 입력 자질을 출력 값으로 분류한다. 예를 들어, 이미지 분류에서 널리 사용되는 CNN의 아키텍처를 보면 컨볼루션 층과

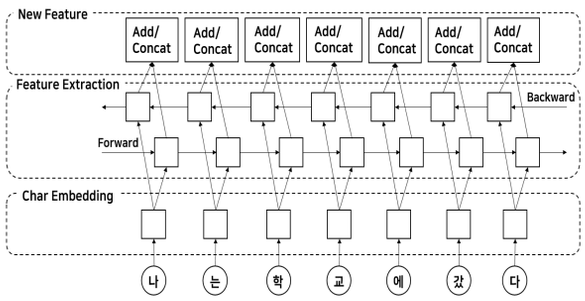

Fig. 3. Bi-LSTM Encoder

풀링 층(pooling layer)의 결합을 통해 이미지 특징을 고정된 길이의 벡터로 표현하고, 이 벡터는 전방향 신경망을 통해 이미지를 분류한다[15]. 본 연구에서는 양방향 LSTM으로 각 현재 음절마다 과거, 미래의 자질 정보와 의존성을 고려하여 생성한 새로운 문맥 자질 벡터와 분류 망으로 널리 사용되는 전방향 신경망을 결합하여 Fig. 4와 같이 한국어 문장의 각 음절의 띄어쓰기 유형을 BI 태그로 분류한다.

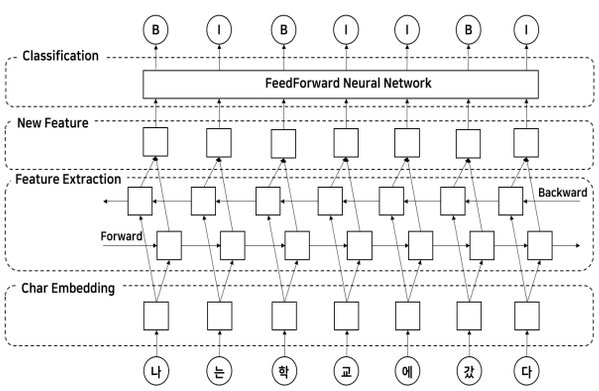

Fig. 4. Automatic Word Spacing of Korean Sentence using Feedforward Neural Network and Bi-LSTM Encoder

3.3 신경망 언어 모델을 이용한 자동 띄어쓰기

언어 모델은 음성 인식, 기계 번역, 정보 검색에서 널리 쓰이는 모델로 일련의 단어 열이 있을 때 다음에 나오는 단어의 확률 분포를 Equation (1)과 같이 예측하는 모델이다[12, 16, 17]. 예를 들어, “나는 학교를”이라는 단어 열이 존재하면 다음에 나올 단어의 확률 분포를 계산하여 “갔다”라는 단어를 예측하는 모델이다.

$$P(w_1,\ldots,w_N)=\prod_{i=1}^{N}P(w_i|w_1,\ldots,w_{i-1}) \quad (1)$$

본 논문에서는 Equation (1)과 같이 결합확률(joint probability)의 연쇄 법칙(chain rule)으로 문장을 구성하는 단어 열의 확률 분포를 계산하는 단방향 LSTM을 이용하는 언어 모델의 방식을 BI 태그를 분류하는 방식으로 변경하여 Fig. 5와 같이 문장의 각 음절을 자동 띄어쓰기 태그로 분류한다[12, 16, 18,

19]. 양방향 LSTM을 이용하여 현재 음절에 과거와 미래의 정보를 고려하여 표현한 문맥 자질 벡터를 신경망 언어 모델의 입력으로 하고 문장의 각 음절에 대한 분류 태그에 대한 확률 분포를 계산하고 한국어 문장의 각 음절의 띄어쓰기 유형을 BI 태그로 분류한다.

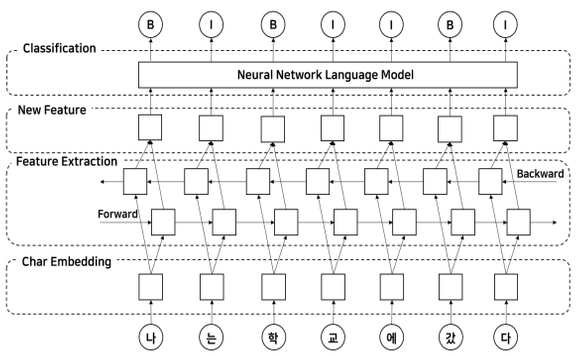

Fig. 5. Automatic Word Spacing of Korean Sentence using Neural Network Language Model and Bi-LSTM Encoder

3.4 선형체인 CRF를 이용한 자동 띄어쓰기

선형체인 CRF는 순차적인 태그 열 분류에서 각 열의 분류 태그만 고려하지 않고 각 열의 분류 태그와 최적의 분류 태그 열을 함께 고려한다[13,14]. 예를 들어, Fig. 6와 같이 "나는학교에갔다"라는 공백이 제거된 한국어 문장의 각 음절의 유형을 띄어쓰기 태그로 분류하면 "BIBIIBI"로 분류한다. 이때, 선형체인 CRF는 각 음절의 띄어쓰기 태그가 B인지 I인지 분류하는 지역적 정보와 문장의 띄어쓰기 태그 열이 "BIBIIBI"인지 확인하는 글로벌 정보를 이용하여 분류 태그 가능성을 계산한다. 본 연구에서는 양방향 LSTM을 이용하여 각 음절에 과거와 미래의 의존성을 고려하여 새로운 문맥 자질 벡터를 표현한다. 이를 각 음절의 B 또는 I 태그에 대한 지역적 점수와 각 음절의 태그와 이웃하는 태그와 의존성을 계산하여 최적의 태그 열인지 아닌지를 선형체인 CRF를 이용하여 로그 가능도(log likelihood)를 계산하고 한국어 문장의 각 음절의 띄어쓰기 유형을 BI 태그로 분류한다.

4. 실험 및 평가

4.1 학습 데이터 구성 및 실험 방법

자동 띄어쓰기의 학습 및 평가를 위한 말뭉치로 최근에 공개된 한국어 원시 말뭉치인 KCC150을 사용하였다.[1] 자동 띄어쓰기 학습 및 테스트를 위한 데이터는 Table 1과 같이 한 라인에 하나의 문장으로 구성하였고, 총 6,981,843개의 문장(어절수: 90,016,390개, 음절수: 286,123,621개) 데이터를 구축하였다. 자동 띄어쓰기의 학습 및 성능 평가를 위해 6,283,656

1) KCC150은 1억 8천만 어절만(약 1,330만 문장)의 규모로 http://nlp.kookmin.ac.kr/kcc에서 다운로드가 가능하다.

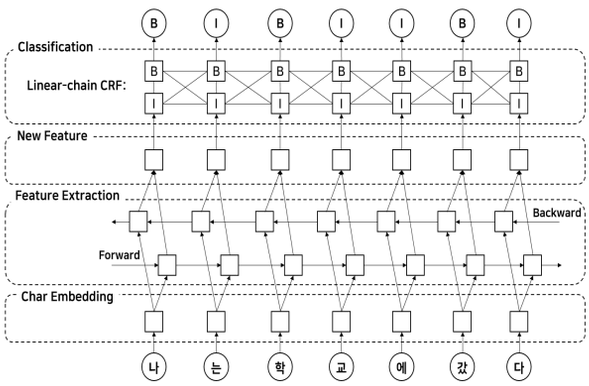

Fig. 6. Automatic Word Spacing of Korean Sentence using Linear-chain CRF Model and Bi-LSTM Encoder

개의 문장(어절수: 81,014,779개, 음절수: 257,516,828개)의 학습 데이터 집합과 698,187개의 문장(어절수: 9,001,611개, 음절수: 28,606,793개)의 테스트 데이터 집합으로 구성하였다.

Table 1. Training and Test Set for Automatic Word Spacing of Korean Sentence

	Training	Test	Total
Lines	6,283,656	698,187	6,981,843
Words	81,014,779	9,001,611	90,016,390
Syllables	257,516,828	28,606,793	286,123,621

분류 망에 따른 자동 띄어쓰기 성능 평가를 위하여 문장 데이터는 전처리 과정으로 "|()[]{}" 등과 같은 특수기호 등은 제거하고 깨끗한 문장 형태의 데이터를 기반으로 학습 및 평가를 수행하였다. 종단 간 심층 신경망 모델의 학습을 위해 배치 사이즈는 1로 하고, 확률적인 경사 하강법(stochastic gradient descent)을 이용한 역전파(backpropagation)로 학습하였다. 그리고 학습률(learning rate)은 0.001, 학습 횟수(epoch)는 1, 5로 학습하였다.

전방향 신경망, 신경망 언어 모델, 그리고 선형체인 CRF의 분류 망에 따른 세 가지 모델들은 자동 띄어쓰기 태그 분류 성능 측정을 위해 음절 임베딩 크기는 300, 새로운 자질 벡터 생성을 위한 양방향 LSTM 셀의 유닛(unit) 크기는 200으로 하고 양방향 LSTM의 정방향 LSTM 셀과 역방향 LSTM 셀의 출력을 더하거나(add), 이어 붙이는(concatenation) 형태의 두 가지의 연산으로 양방향 LSTM의 출력을 나누어 새로운 문맥 자질 벡터를 연속적인 벡터 공간에 표현하였다. 그리고 전방향 신경망의 경우에는 은닉층(hidden layer) 유닛 크기로 200, 활성화 함수(activation function)는 ReLU로 층수를 1에서 6까지 한 층씩 늘리면서 자동 띄어쓰기 성능 평가를 하였고, 단방향 LSTM을 이용한 신경망 언어 모델 경우 단방향 LSTM 셀의 유닛 수를 200으로 하고 단방향 LSTM을 단층으로 하여 성능 평가를 하였다. 전방향 신경망과 단방향

LSTM을 이용한 신경망 언어 모델은 소프트맥스(softmax)와 교차 엔트로피(cross entropy)를 이용하여 문장의 각 음절의 분류 태그의 확률 분포를 계산하여 띄어쓰기 BI 태그로 분류하였다. 그리고 선형체인 CRF 모델의 경우 지역적 점수 계산을 위해 출력 층 하나만 존재하는 전방향 신경망을 이용하여 지역적 점수를 계산하였다.

4.2 실험 결과 및 성능 평가

모델의 정량적 성능 평가를 위해 Equation (2)-(4)과 같이 자동 띄어쓰기 태그 분류 정확도(accuracy), 공백 재현율(spacing recall), 공백 정확도(spacing precision)를 사용하였다. Equation (2)는 (올바르게 예측된 BI 태그 수)/(실제 전체의 BI 태그 수)*100으로 자동 띄어쓰기 태그 분류 정확도이고, Equation (3)은 공백 삽입 관점에서 (올바르게 예측된 공백 위치 수)/(실제 전체의 공백 위치 수)*100으로 공백 재현율을 나타낸다. Equation (4)는 (예측된 올바른 공백 수)/(예측된 전체 공백 수)*100으로 공백 정확도이다. Equation (3)과 Equation (4)를 이용하여 F1 값을 계산하고 성능 평가를 하였다.

$$Accuracy = \frac{the\,predicted\,correct\,tags}{the\,actual\,entire\,tags} \times 100 \quad (2)$$

$$Spacing\,Recall = \frac{the\,predicted\,correct\,spacing}{the\,actual\,entire\,spacing} \times 100 \quad (3)$$

$$Spacing\,Precision = \frac{the\,predicted\,correct\,spacing}{the\,predicted\,entire\,spacing} \times 100 \quad (4)$$

Table 2는 전방향 신경망을 이용한 실험 결과이다. 전방향 신경망의 경우에는 층수에 따른 자동 띄어쓰기 태그 분류 정확도, 공백 재현율 그리고 공백 정확도에 대한 성능 평가 결과를 보면 층수가 깊어짐에 따라 자동 띄어쓰기 태그 정확도와 공백 재현율의 성능이 개선되는 효과를 보여주었고, 자동 띄어쓰기 태그 정확도 97.14%, 공백 재현율 95.96%, 공백 정확도 95.01%, F1 값은 95.10%를 보여주었다.

Table 3은 전방향 신경망, 신경망 언어 모델, 그리고 선형체인 CRF의 분류 망에 따른 종단 간 심층 신경망 모델들의 성능을 평가한 결과이다. 한국어 문장의 자동 띄어쓰기 성능 평가 결과는 각 분류 방법론에 따른 태그 분류 정확도, 공백 재현율, 공백 정확도, F1 값에서는 선형 체인 CRF 분류 방법을 사용했을 때, 자동 띄어쓰기 태그 정확도 97.91%, 공백 재현율 96.49%, 공백 정확도 96.31%, F1 값은 96.40%를 보여주었다.

Table 2. The Performance of Feedforward Neural Network

Bi-LSTM Output	Layer	Accuracy	Spacing Recall	Spacing Precision	F1 Score
Add	1	96.19	92.66	94.13	93.39
	2	96.73	93.80	94.87	94.33
	3	97.05	94.84	95.01	94.92
	4	97.12	95.12	94.97	95.04
	5	97.10	95.71	94.39	95.05
	6	97.13	95.37	94.77	95.07
Concatenation	1	96.13	92.53	94.03	93.27
	2	96.70	93.71	94.86	94.28
	3	96.97	95.30	94.30	94.80
	4	96.99	95.28	94.39	94.83
	5	97.13	95.96	94.25	95.10
	6	97.14	95.62	94.58	95.10

Table 3. The Performance of Automatic Word Spacing of Korean Sentences Depending on Classification Network

Classification Neural Network	Bi-LSTM Output	Accuracy	Spacing Recall	Spacing Precision	F1 Score
Feedforward Neural Network	Add	97.13	95.37	94.77	95.07
	Concatenation	97.14	95.62	94.58	95.10
Neural Network Language Model	Add	96.77	94.98	93.96	94.47
	Concatenation	96.70	95.04	93.68	94.36
Linear-chain CRF	Add	97.91	96.49	96.31	96.40
	Concatenation	97.68	96.29	95.75	96.02

Table 4. The Performance of the Existing Model on Korean Automatic Word Spacing

Model	Syllable Accuracy	Eojeol Precision	Eojeol Recall	F1 Score	Data
Shim(2015)	98.06	92.27	94.15	93.20	Sejong Corpus
Hwang(2016)	98.32	92.68	91.96	92.32	Sejong and ETRI Corpus
Lee(2018)	-	93.72	94.27	93.99	Sejong Corpus
Jeon(2018)	97.1	-	-	-	Sejong Corpus
	94.3	-	-	-	Literary Style Corpus

4.3 기존 연구와 비교 평가

Table 4는 기존 연구의 실험 결과이다. 기존의 연구에서 [1]은 어절 빈도 사전과 음절 빈도 사전 기반의 어절 정보와 형태소 분석기를 이용하였고, 세종 말뭉치에서 순수 한글 585만 어절을 발췌하여 10 배수 교차 검증으로 실험한 결과 98.06%의 음절 정확도와 94.15%의 어절 재현율을 얻었다. 심층 신경망을 이용한 자동 띄어쓰기 방법으로, [5]는 GRU-CRF 모델은 입력 자질로 음절 unigram, bigram, trigram의 조합과 명사 사전을 이용하여 과적합(overfitting) 문제를 해결하기 위하여 입력층과 은닉층에 드롭아웃(dropout) 기술을 적용하였다. 그리고 어절 26,013,702개의 세종 말뭉치의 학습 데이터와 288,291개 어절 ETRI 말뭉치를 평가 데이터로 이용하여 학습 및 성능을 평가한 결과 92.32%의 F1 점수를 보여주었다.

[6]은 LSTM 기반의 Sequence-to-Sequence 모델을 각각 층수가 3 층인 인코더와 디코더로 구성하였다. 인코더는 정방향 LSTM 2층과 역방향 1층으로 구성하고 디코더는 LSTM 3층으로 구성하여 드롭아웃, 계층 정규화, 주목 기법(attention mechanism)의 기술을 적용하였다. 그리고 한 라인에 최대 단어 10개로 구성하고, 총 6,161,374의 행의 세종 말뭉치를 이용하여 학습 및 성능 평가를 한 결과 94.0%의 F1 점수를 얻었다. 또한, [7]은 n-gram의 인코딩을 위한 1차원 컨볼루션과 GRU으로 구성된 모델은 1M 크기의 테스트 말뭉치인 세종 말뭉치에서는 97.1%, 3M 크기의 테스트 말뭉치인 문학 스타일의 말뭉치에서는 94.3%의 음절 정확도를 보여주었다.

[1]은 585만 어절, [5]는 288,291개 어절의 평가 데이터, [7]의 1M 크기의 데이터보다 많은 대용량 사이즈의 총 90,016,390개 어절에서 9,001,611개 어절의 평가 데이터를 사용하여 성능을 평가한 결과 기존의 모델과 필적하거나 우수한 성능을 보여주었다. 또한, [7]에서 ngram의 인코딩을 위한 1차원 컨볼루션과 GRU으로 구성된 모델은 세종 말뭉치에서는 97.1%, 문학 스타일의 말뭉치에서는 94.3%의 음절 정확도를 보여주었고, 본 논문에서는 음절 unigram만 이용하여 양방향 LSTM으로 인코딩하고 분류 망을 전방향 신경망과 선형체인 CRF로 한 경우에 각각 97.14%, 97.91%의 음절 정확도를 보여주어 [7]보다 우수한 성능을 보여주었다. 신경망 언어 모델을 분류 망으로 하는 경우에는 96.77%의 음절 정확도를 보여주어 3M 크기의 문학 스타일의 말뭉치에서 평가한 [7]보다 우수한 성능을 보여주었다.

기존의 연구에서 공통적으로 사용한 말뭉치는 세종 말뭉치로 본 연구에서 제안한 다양한 분류 방법에 따른 모델 중 우수한 성능을 보여주는 선형 체인 CRF의 분류 방법 중 정방향 LSTM과 역방향 LSTM의 출력값을 더하는 모델로 세종 말뭉치를 학습하고 실험하였다. 전처리 과정으로 “|()[]{}” 등과 같은 특수기호 등은 제거하고 깨끗한 3,730,158개의 문장 데이터를 3,357,143개의 문장(어절수: 46,480,665개, 음절수: 141,686,680개)의 학습 데이터 집합과 373,015개의 문장(어절수: 560,912개, 음절수: 15,734,223개)의 테스트 데이터 집합으로 구성하였다. 하이퍼 파라미터 세팅은 KCC150 데이터로 학습 및 성능 평가 시와 동일하게 세팅을 하였다. 그 결과, 음절 정확도 97.167%, 공백 정확도 95.56%, 공백 재현율 95.11%, 어절 정확도 88.573%, 어절 재현율 88.164%을 보여주었다.

어절 정확도와 어절 재현율이 낮게 평가된 이유는 복합명사와 복합용언의 경우에 띄어쓰기와 붙여쓰기가 모두 허용되는데 원문과 다르면 틀린 것으로 평가했기 때문이다. 복합명사와 복합용언의 띄어쓰기와 관련된 평가 오류를 바로 잡아 재평가를 수행하기 위해 1,000개의 문장을 샘플링한 후, 성능 평가를 한 결과 음절 정확도 98.533%, 공백 정확도 98.49%, 공백 재현율 96.73%, 어절 정확도 95.068%, 어절 재현율 93.468%이고 어절 F1 94.261%의 성능을 보여주었다. 이는 어절 성능 측면에서 기존의 연구에서 가장 우수한 성능을 보였던 [6]보다 F1 0.271%만큼 향상되었다. 그리고 음절 정확도에서는 기존의 연구에서 우수한 [5]보다 음절 정확도에서 0.213%만큼의 향상된 성능을 보여주었다.

5. 결 론

어절 단위가 아닌 음절 단위로 한국어 문장 자동 띄어쓰기 문제를 추가적인 형태소 분석 없이 종단 간 심층 신경망 시스템으로 구성하였다. 또한, 자동 띄어쓰기 문제를 음절 태그 분류 문제로 정의하여 음절 임베딩과 양방향 LSTM을 이용하여 고정된 길이의 새로운 자질 벡터를 연속적인 벡터 공간에 표현하고 심층 신경망 모델에서 분류 망으로 널리 사용되는 전방향 신경망, 신경망 언어 모델, 그리고 선형 체인 CRF

로 분류 망이 다른 세 가지 심층 신경망 모델로 한국어 문장에 대한 자동 띄어쓰기 성능 평가를 하였다. 성능 평가 결과 같은 하이퍼 파라미터(hyper parameter)에서 분류 망에 따른 세 개의 모델 중에서 선형 체인 CRF가 우수함을 보여주었고, 전방향 신경망의 경우 층수가 깊어짐에 따라 자동 띄어쓰기 성능이 개선되어 단방향 LSTM을 이용한 신경망 언어 모델보다 우수함을 보여주었다.

References

[1] K. S. Shim, "Automatic Word Spacing using Raw Corpus and a Morphological Analyzer," *Journal of KIISE*, Vol.42, No.1, pp.68-75, 2015.

[2] K. S. Kim, H. J. Lee, and S. J. Lee, "Three-stage Word-spacing System for Continuous Syllable Sentence in Korea," *Journal of KISS(B): Software and Applications*, Vol.25, No.12, pp.1838-1844, 1998.

[3] S. S. Kang, "Eojeol-block Bidirectional Algorithm for Automatic Word Spacing of Hangul Sentences," *Journal of KISS : Software and Applications*, Vol.27, No.4, pp.441-447, 2000.

[4] C. K. Lee, "Structural SVM-based Korean Word Spacing using Spacing Information Input by Users," *Journal of KIISE: Computing Practices and Letters*, Vol.20, No.5, pp.301-305, 2014.

[5] H. S. Hwang and C. K. Lee, "Automatic Korean Word Spacing using Deep Learning," in Korea Computer Congress of KIISE, Jeju, The South Korea, 2016, pp.738-740.

[6] T. S. Lee and S. S. Kang, "LSTM Based Sequence-to-sequence Model for Korean Automatic Word-spacing," *Smart Media Journal*, Vol.7, No.4, pp.17-23, 2018.

[7] Heewon Jeon, "KoSpacing: Automatic Korean Word Spacing," GitHub Repository, https://github.com/haven-jeon/PyKoSpacing, http://freesearch.pe.kr/archives/4759, 2018.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," in *Advances in Neural Information Processing Systems*, Lake Tahoe, the United States, 2013, pp.3111-3119.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv Preprint arXiv:1301.3781, 2013.

[10] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, Vol 5, pp.135-146, 2017.

[11] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp.1532-1543.

[12] T. Mikolov, M. Karafiát, L. Burget, J. Černocký and S. Khudanpur, "Recurrent Neural Network Based Language Model," in Eleventh Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, 2010, pp.1045-1048.

[13] S. W. Kim, and S. P. Choi, "Research on Joint Models for Korean Word Spacing and POS (Part-Of-Speech) Tagging Based on Bidirectional LSTM-CRF," *Journal of KIISE*, Vol.45, No.8, pp.792-800, Aug, 2018.

[14] Z. H. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," arXiv Preprint arXiv: 1508.01991, 2015.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems. Harrahs and Harveys, Lake Tahoe, the United States, 2012, pp.1097-1105.

[16] T. Mikolov, S. Kombrink, L. Burget, J. Černocký, and S. Khudanpur, "Extensions of Recurrent Neural Network Language Model," in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 2011, pp.5528-5531.

[17] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, Vol 3, pp.1137-1155, 2003.

[18] M. Sundermeyer, R. Schluter, and H. Ney, "LSTM Neural Networks for Language Modeling," in Thirteenth Annual Conference of the International Speech Communication Association, Portland, OR, USA, 2012, pp.194-197.

[19] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the Limits of Language Modeling," arXiv Preprint arXiv:1602.02410, 2016.

이 현 영

https://orcid.org/0000-0003-2553-6576
e-mail : le32146@gmail.com
2016년 국민대학교 컴퓨터공학부(학사)
2016년~2017년 SK Hynix Memory Solutions Inc. Intern
2019년 국민대학교 컴퓨터공학과(석사)
2019년~현 재 국민대학교 컴퓨터공학과 박사과정
관심분야 : 자연어처리, 기계학습, 인공지능, 정보검색, 빅데이터 분석, 추천시스템

강 승 식

https://orcid.org/0000-0003-3318-6326

e-mail : sskang@kookmin.ac.kr

1986년 서울대학교 컴퓨터공학과(학사)

1988년 서울대학교 컴퓨터공학과(석사)

1993년 서울대학교 컴퓨터공학과(박사)

2001년～현 재 국민대학교 소프트웨어학부 교수

관심분야 : 자연어처리, 텍스트마이닝, 빅데이터 분석, 상황인지 컴퓨팅