KIPS Tr. Software and Data Eng.
Vol.7, No.4 pp.121~128 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2018.7.4.121

A Reranking Model for Korean Morphological Analysis Based on Sequence-to-Sequence Model

Yong-Seok Choi† · Kong Joo Lee††

ABSTRACT

A Korean morphological analyzer adopts sequence-to-sequence (seq2seq) model, which can generate an output sequence of different length from an input. In general, a seq2seq based Korean morphological analyzer takes a syllable-unit based sequence as an input, and output a syllable-unit based sequence. Syllable-based morphological analysis has the advantage that unknown words can be easily handled, but has the disadvantages that morpheme-based information is ignored. In this paper, we propose a reranking model as a post-processor of seq2seq model that can improve the accuracy of morphological analysis. The seq2seq based morphological analyzer can generate K results by using a beam-search method. The reranking model exploits morpheme-unit embedding information as well as n-gram of morphemes in order to reorder K results. The experimental results show that the reranking model can improve 1.17% F1 score comparing with the original seq2seq model.

Keywords : Sequence-to-Sequence Model, Reranking, Beam-Search, Syllable-Unit Processing

Sequence-to-Sequence 모델 기반으로 한 한국어 형태소 분석의 재순위화 모델

최 용 석† · 이 공 주††

요 약

Sequence-to-sequence(Seq2seq) 모델은 입력열과 출력열의 길이가 다를 경우에도 적용할 수 있는 모델로 한국어 형태소 분석에서 많이 사용되고 있다. 일반적으로 Seq2seq 모델을 이용한 한국어 형태소 분석에서는 원문을 음절 단위로 처리하고 형태소와 품사를 음절 단위로 출력한다. 음절 단위의 형태소 분석은 사전 미등록어 문제를 쉽게 처리할 수 있다는 장점이 있는 반면 형태소 단위의 사전 정보를 반영하지 못한다는 단점이 있다. 본 연구에서는 Seq2seq 모델의 후처리로 재순위화 모델을 추가하여 형태소 분석의 최종 성능을 향상시킬 수 있는 모델을 제안한다. Seq2seq 모델에 빔 서치를 적용하여 K개 형태소 분석 결과를 생성하고 이들 결과의 순위를 재조정하는 재순위화 모델을 적용한다. 재순위화 모델은 기존의 음절 단위 처리에서 반영하지 못했던 형태소 단위의 임베딩 정보와 n-gram 문맥 정보를 활용한다. 제안한 재순위화 모델은 기존 Seq2seq 모델에 비해 약 1.17%의 F1 점수가 향상되었다.

키워드 : Sequence-to-Sequence 모델, 재순위화, 빔 서치, 음절 단위 처리

1. 서 론

최근 딥러닝 모델이 발표된 이후 한국어 형태소 분석 및 품사 부착 문제에서도 딥러닝을 적용한 연구들이 좋은 성능을 보이고 있다[1-3]. 딥러닝 모델을 한국어 형태소 분석에 적용하면서 입력 문장(원문)으로부터 그에 대한 형태소 분석 결과를 바로 생성하는 연구가 주목받고 있다. 그 중 Sequence-to-sequence(Seq2seq)모델은 다양한 길이의 입력열과 출력열을 표현할 수 있는 모델로 한국어 형태소 분석에서도 많이 사용되고 있다.

[2]와 [3]의 연구는 Seq2seq를 적용하여 원문을 음절 단위의 입력으로 받아 형태소와 품사 부착이 이루어진 음절 단위의 결과열을 출력한다. 형태소 분석기의 입력과 출력을 음절 단위로 사용한 이유는 형태소 분석 과정 중 발생하는 사전 미등록어(Out of Vocabulary; OOV) 문제를 효율적으로 처리할 수 있기 때문이다. 그렇지만 음절 단위로 처리를 할 경우 사전 정보를 반영하지 못하거나 원형 복원을 제대로 하지 못하는 문제가 발생할 수 있다[4].

※ 이 논문은 2017년 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017M3C4A7068187).
† 준 회 원 : 충남대학교 전자전파정보통신공학과 박사과정
†† 종신회원 : 충남대학교 전파정보통신공학과 교수
Manuscript Received : March 15, 2018
Accepted : March 27, 2018
* Corresponding Author : Kong Joo Lee(kjoolee@cnu.ac.kr)

다음 (예 1)은 실제 seq2seq 모델 기반의 한국어 형태소 분석기가 출력한 결과 중 오류의 일부이다

(예 1)

(1) “참다못한”: 참/VV 다못/EC 하/VX ㄴ/ETM
(2) “두꺼웠다고”: 두꺼우/VV 었/EP 다고/EC
(3) “버티어” : 버쉽/VV 어/EC
(4) “높낮이도”: 높낮/NNG 이도/JX

본 연구에서는 Seq2seq 모델 기반의 한국어 형태소 분석기가 가지는 오류에 대해 이를 해결하기 위해 재순위화(Reranking) 모델을 제안한다. 일반적으로 Seq2seq 모델 기반의 형태소 분석기는 하나의 형태소 분석열을 결과로 생성한다. 본 연구에서는 [3]의 모델로부터 단일 출력열 대신 K개의 형태소 분석 후보 결과열을 생성(K-best)하고 그 후보 결과열들의 순위를 다시 재조정하여 최종 형태소 분석 결과열을 결정하는 모델을 제안한다. K개의 형태소 분석 후보 결과열은 [5]의 연구에서 제안한 Seq2seq 모델의 디코더 단계에서 빔 서치(Beam Search) 알고리즘으로 생성한다. 생성한 K개의 후보 결과열에 대해 Seq2seq 모델에서 다루지 못한 형태소 단위의 정보를 활용하여 재순위화 한다.

Seq2seq 모델을 이용한 한국어 형태소 분석기[2, 3]는 입출력 모두 음절 단위이다. 그렇기 때문에 최종 형태소 분석 결과는 출력열에 나타나는 품사 정보 기준으로 앞의 나온 형태소 음절들을 해당 품사의 형태소로 간주하여 품사를 부착한다. 본 연구에서는 형태소 분석 결과의 순위를 재조정하기 위해 출력 음절들을 형태소 단위로 재구성한 후, 형태소 단위의 n-gram 정보를 이용하여 형태소 분석 결과에 점수를 부여하고 이 점수를 이용하여 재순위화를 수행한다.

음절열로 출력된 결과로부터 형태소 단위의 재구성을 위해서는 각각의 음절 정보와 음절열로 구성된 형태소의 단어 정보, 그리고 해당 형태소의 품사 정보를 모두 임베딩 벡터로 표현하고 이를 결합하여 하나의 형태소 정보로 재구성하였다.

이와 같이 형태소 단위로 재구성된 형태소 분석열에서 컨볼루션 신경망(Convolutional Neural Networks) 모델을 사용하여 형태소 단위의 문맥(n-gram) 정보를 학습시키고, 이를 점수화하였다. 이 점수를 통해 후보 형태소 분석열들의 순위를 재조정 한 후, 가장 높은 점수를 받은 형태소 분석열을 최종 분석 결과로 결정한다.

본 논문의 구성은 다음과 같다. 2장에서 Seq2seq 모델, 재순위화 그리고 단어를 표현하는 방법과 관련된 연구들을 살펴본다. 3장에서는 본 연구에서 제안한 모델을 설명한다. 4장에서는 제안한 모델의 실험 및 결과 분석으로 구성된다. 5장에서는 결론으로 논문 구성을 마친다.

2. 관련 연구

2.1 Seq2seq 모델 기반 한국어 형태소 분석기

Seq2seq 모델 기반 한국어 형태소 분석기는 한국어 형태소 분석 문제를 입력 원문에서 품사 정보가 부착된 형태소 분석 결과로 생성하는 것으로 번역하는 문제로 간주하여 적용을 한다. Seq2seq 모델은 원문 문장을 하나의 벡터로 압축하여 표현해준 인코더와 인코더에서 표현된 벡터로부터 형태소와 품사를 생성하는 디코더로 나눌 수 있다. [2, 3]의 연구는 Seq2seq 모델을 이용하여 한국어 형태소 분석 문제를 해결하였으며, 형태소의 미등록 단어 문제(OOV)를 해결하기 위해 입력과 출력 모두 음절 단위로 구성한다.

Seq2seq 모델의 인코더와 디코더는 하나의 재귀 신경망(Recurrent Neural Network)으로 구성되어 있다. 일반적으로 재귀 신경망 모델은 입력열이 길어질 경우 학습 과정에서 기울기의 값이 사라지는 문제(Vanishing Gradient Problem)가 있다. 이를 해결하기 위해 LSTM(Long Short-Term Memory)와 GRU(Gated Recurrent Unit)들이 있다.

Seq2seq 모델은 입력열에 문장이 길어지면 인코더에서 표현해야하는 정보량이 커져 모델의 성능이 떨어지는 문제가 있다. Seq2seq 모델을 사용하는 한국어 형태소 분석기도 이러한 문제를 가지고 있다. [2]의 연구에서는 주의 기반(Attention-based) 알고리즘을 사용하여 이 문제를 해결하고자 하였다. 주의 기반 알고리즘은 디코더의 i번째 단계를 계산할 때, 입력의 어느 부분을 더 반영할 것인가를 가중치로 표현해준다. 인코더의 은닉 변수들($h_1, h_2, \cdots, h_{j-1}, h_j, h_{j+1}, \cdots, h_{|x|}$)과 디코더의 i-1번째 은닉 변수 S_{i-1}를 신경망의 입력으로 하여 주의 기반 가중치(Attention Weights) a_{ij}를 계산한다.

$$C_i = \sum_{j=1}^{|x|} a_{ij} h_j \tag{1}$$

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{|x|} \exp(e_{ik})} \tag{2}$$

$$e_{ij} = a(s_{i-1}, h_j) \tag{3}$$

$$S_i = f(S_{i-1}, y_{i-1}, C_i) \tag{4}$$

[2]에서 제안한 주의기반 알고리즘을 이용한 Seq2seq모델 기반 한국어 형태소 분석기의 경우는 외래어와 같이 고유 명사에만 주로 사용되는 빈도수 낮은 음절들이 디코더 단계에서 등장하지 않는 문제가 발생할 수 있다. 예를 들어 입력열 ‘삿포로’[3]에 대해 ‘삿’은 학습 데이터에 많이 발생하지 않는 음절이기 때문에 형태소 분석 출력열에 다른 음절로 바뀌어 나타나는 경향이 있다.

이 문제를 해결하기 위해 [3]의 연구에서는 복사 작용(Copying Mechanism) 알고리즘을 이용한 Seq2seq 모델을 사용한다. 복사 작용 알고리즘[6]은 디코더 단계에서 인코더 단계의 입력열에 포함된 음절의 확률을 합하여 최종 확률 값을 계산한다. 이 과정을 통해 입력열에 존재하는 음절들이 디코더 단계에서도 음절들이 나타날 확률을 높여줌으로 빈도수가 낮은 음절들이 분석 결과에서 배제되는 문제를 해결할 수 있다.

본 연구에서는 한국어 형태소 분석 연구에 가장 최근 방법인 [3]의 연구를 기본적으로 활용하고 그 분석 결과에 대해 후처리를 통하여 순위를 재조정해보고자 한다.

2.2 Seq2seq 모델 결과를 이용한 재순위화

Seq2seq는 종단 간 모델로 하나의 입력열을 통해 하나의 출력열을 생성한다. 디코더 단계에서 각 노드의 출력을 결정할 때는 이전 시퀀스 정보만 반영된다. 그렇기 때문에 각 노드마다 하나의 출력만을 가지고 진행하므로 한 번 오류가 발생하면 그대로 누적된 오류 결과가 발생한다. 이 문제를 해결하기 위해 빔 서치 알고리즘이 Seq2seq 모델에 적용되었다[5, 7]. 빔 서치 알고리즘은 최고 우선 탐색(Best-First Search) 방법으로 t번째 단계에서 t+1번째 단계에 선택될 수 있는 가능한 모든 경우의 수를 계산한 후 상위 K개의 결과만 다음 단계의 출력 결과로 내보낸다.

Seq2seq 모델은 입력과 출력만을 사용하기 때문에 도움이 되는 정보가 있어도 이를 모델에 직접 적용하는 것이 어렵다. 그렇기 때문에 Seq2seq 모델에 빔 서치 알고리즘을 이용하여 출력한 결과에 대해 다른 정보를 활용하여 순위를 재조정하는 연구들이 진행되고 있다[8-11].

[8]의 연구는 Seq2seq 모델의 입력 x와 출력 y 사이의 상호 정보(mutual information)를 사용하기 위해 Seq2seq 모델의 출력 결과에 대해 순위를 재조정하는 연구를 시도하였다. 입력을 통해 출력이 생성하는 $Seq2seq_{(x\rightarrow y)}$ 모델과 출력에서부터 입력이 생성되는 $Seq2seq_{(y\rightarrow x)}$ 모델 2개를 구축한다. 첫 번째 $Seq2seq_{(x\rightarrow y)}$ 모델로부터 입력 x에 대해 빔 서치 알고리즘을 사용하여 K개의 출력을 생성한다. K개의 출력에 대해 $Seq2seq_{(y\rightarrow x)}$ 모델을 활용하여 순위를 재조정하고 이를 통해 최종 결과를 얻어낸다. 이 외에도 추가 정보를 활용하기 위해 출력 y의 언어 모델도 사용하였다. [8] 연구의 재순위화를 위한 점수 계산은 Equation (5)와 같다.

$$Score(y) = \log p(y|x) + \lambda \log p(x|y) + \gamma \log p(y) + \eta LT \qquad (5)$$

log p(y)는 출력에 대한 언어 모델이고 LT는 출력 문장의 길이이이다. 이 수식을 영어-독일어 번역에 사용하였을 때, 기존 모델 대비 약 2점 이상의 BLEU 점수가 향상되는 것을 확인할 수 있었다.

[9]의 연구에서는 자연 언어 생성(NLG) 분야에 Seq2seq 모델과 출력 결과를 통해 순위를 재조정하였다. 기존의 자연 언어 생성은 문장 계획(Sentence Planning)과 표층문 실현(Surface Realization)으로 2단계로 나누어 구현하였다. 반면, [9]의 연구에서는 이 두 단계를 종단 간 모델로 구현하기 위해 Seq2seq를 사용하였다. 이 때, 입력은 화행(dialogue act)이고 출력은 표층 문장(surface sentence)이다. 이 연구에서는 출력 문장이 입력의 화행과 의미적으로 일치하는지 확인하기 위해 Seq2seq 모델에서 빔 서치 알고리즘을 사용하여 K개의 결과를 얻어낸다.

재순위화 방법으로는 출력 문장 대해 각 화행 종류와 해당 슬롯 값의 존재 여부를 [12]의 분류기를 통해 얻어서 각각을 바이너리(1-hot) 벡터로 표현하는 방법이다. 입력 문장 또한 바이너리 벡터로 표현한 후 두 벡터의 거리 값을 이용하여 K개의 출력에 대해 순위를 재조정하였다. 출력 문장을 생성하기 위해 빔의 크기를 100개로 늘려서 순위를 재조정할 경우 BLEU 점수 값을 52.54에서 62.76까지 향상 되는 것을 확인할 수 있었다.

2.3 음절 기반의 단어 임베딩 표현

단어의 표현은 신경망 모델이 발표된 이후 임베딩 벡터로 표현하며 자연 언어 처리 분야에서 기초적인 입력으로 사용되어지고 있다[13]. 단어 임베딩 벡터의 가장 큰 장점은 복잡한 자질 설계를 할 필요가 없으며 신경망을 통해 단어의 구문 정보와 의미 정보를 학습시킬 수 있다는 것이다.

단어 임베딩 표현은 주변 단어들에 의해 구문 정보와 의미 정보를 포함할 수는 있지만 단어의 형태 정보는 얻기가 어려웠으며, 미등록 단어들에 대한 처리도 어렵다는 문제가 있다. 이를 해결하기 위해 단어 임베딩 표현을 글자 단위(또는 음절 단위)로 나누어 조합하는 연구들이 진행되었다[14-16].

[14]의 연구에서는 품사 부착 문제에서 미등록어를 해결하기 위해 음절 기반의 단어 임베딩 벡터를 제안하였다. 음절 기반의 단어 임베딩 벡터는 단어를 이루고 있는 문자들에 컨볼루션 신경망을 취하여 하나의 벡터로 표현을 하였다. 이후 단어 임베딩 벡터와 음절 기반의 단어 임베딩 벡터를 조합하여 품사 부착을 해결해보고자 하였다. 이를 영어와 포르투갈어에 품사 부착에 적용하였을 때, 약 1.3%의 성능 향상을 확인할 수 있었다.

[15]의 연구는 음절 기반의 임베딩 벡터를 개체명 인식 문제에 적용하였다. 이 또한 미등록어 문제를 해결하기 위해 단어를 이루고 있는 문자들을 컨볼루션 신경망을 통해 하나의 단어 벡터로 표현하였다. 기본적인 개체명 인식 모델에 단어 임베딩 벡터와 음절 기반의 단어 임베딩 벡터를 입력으로 사용하였다. 실험 결과 단어 임베딩 벡터만 사용했을 때에 비해 약 0.26%의 성능 향상을 보였다.

3. 형태소 분석 결과열에 대한 재순위화 모델

본 연구에서는 Seq2seq 모델 기반 한국어 형태소 분석이 가지고 있는 오류를 해결해보기 위해 재순위화 모델을 제안한다. 재순위화를 위해서는 다양한 후보 형태소 분석열이 필요하다. 다음 (예 2)는 원문 “하얀 돌을 주웠다”을 입력으로 한 Seq2seq 모델의 후보 형태소 분석열 결과 예제이다. Seq2seq 모델에서 가장 높은 점수를 받아 출력된 결과는 (1)이다. 하지만 (1)의 결과는 ‘하얀/VA’이 원형 복원이 제대로 안되어 분석이 잘못된 경우이다. 그런데 빔 서치 결과의 후보 중에 (3)과 같이 형태소 분석이 제대로 된 결과가 있는 것을 확인할 수 있다.

(예 2)

(1) 하얀/VA 돌/NNG 을/JKO 줍/VV 었/EP 다/EF

(2) 하얗/VA ㄴ/ETM 돌/NNG 을/JKO 주/VV 었/EP 다/EF

(3) 하얗/VA ㄴ/ETM 돌/NNG 을/JKO 줍/VV 었/EP 다/EF

(4) 하얀/VA 돌/NNG 을/JKO 주/VV 었/EP 다/EF
(5) 하얗/VA ㄴ/ETM 돌돌/NNG 을/JKO 줍/VV 었/EP 다/EF

이와 같이 빔 서치를 통해 얻은 다양한 후보 형태소 분석열에는 Seq2seq 모델이 놓친 정답이 존재할 수 있다. 본 연구에서는 형태소의 사전 정보와 형태소 간의 문맥 정보를 반영하여 형태소 분석열들의 순위를 재조정한다.

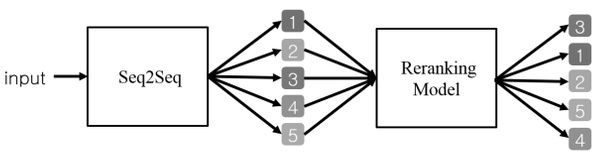

Fig. 1. Reranking Model Concept based on K-best using Seq2seq [4]

본 연구에서는 Fig. 1처럼 Seq2seq 모델의 디코더 단계에서 빔 서치를 통해 K개의 형태소 분석열을 후보로 생성한다. 각 후보 형태소 분석열은 재순위화 모델에 의해 점수를 부여받게 되고 이 점수에 의해 순위를 재조정하게 된다. 재조정된 순위에서 가장 높은 점수를 부여 받은 형태소 분석열이 최종 형태소 분석의 결과이다.

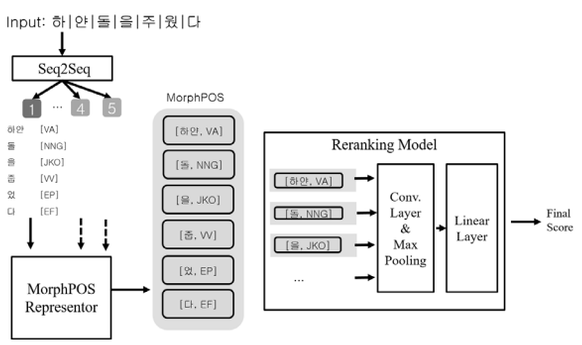

Fig. 2. Overall Processing of Morphological Analysis with Reranking Model

Fig. 2는 재순위화 모델을 포함한 전체적인 흐름도이다. 원문으로부터 음절 단위의 입력을 받아 Seq2seq 모델을 통해 K개의 출력 후보 형태소 분석 결과를 생성한다.

각 형태소 분석 결과에 대해 음절 단위의 분석결과를 형태소 단위의 결과로 재구성하여 하나의 벡터(MorphPOS)로 표현한다. MorphPOS 벡터로 표현된 형태소 분석 결과는 재순위화 모델(Reranking Model)의 입력으로 사용된다. 재순위화에 사용하는 기본 정보는 형태소 단위의 문맥(n-gram) 정보이다. 컨볼루션 신경망(Conv. Layer)과 최대 풀링 레이어(Max Pooling)를 통해 형태소 단위의 문맥 정보가 반영된 벡터를 생성하고 이를 선형 레이어(Linear Layer)를 통해 하나의 점수값을 출력한다. 이 점수 값은 각 후보 형태소 분석 결과의 적합성이다. K개의 출력 후보 형태소 분석 결과를 각각 점수를 계산하고 이를 이용해 순위를 재조정하여 가장 높은 점수를 받은 형태소 분석 결과를 최종 출력으로 결정한다.

3.1 형태소 표현 방법

하나의 형태소는 단어(lexical)와 품사(part-of-speech) 정보를 이용하여 표현한다. 이를 위해 형태소 단어의 임베딩과 품사 정보에 대한 임베딩을 결합하여 하나의 형태소를 표현한다. 형태소 단위로 처리할 때의 문제점 중 하나는 미등록어 처리이다. 본 연구에서는 사전 미등록어를 효율적으로 해결하기 위해 형태소를 형태소 단위의 처리 이외에 음절의 조합으로도 표현한다. 음절 임베딩을 학습시키고 형태소를 구성하는 음절 벡터의 조합으로 형태소를 표현한다. 각 형태소마다 음절의 길이가 다르기 때문에 이를 하나의 고정된 크기의 벡터로 표현해주기 위해 컨볼루션 신경망을 사용한다. 컨볼루션 신경망을 통해 나온 벡터들은 최대 풀링을 통해 하나의 고정된 크기의 벡터로 표현한다. 이를 형태소에 대한 음절 벡터로 정의한다.

Fig. 3은 음절 벡터를 표현하는 방법이다. “세르테르”의 형태소를 음절 벡터로 표현하기 위해 ‘세’, ‘르’, ‘테’, ‘르’의 각 음절 임베딩 벡터($e_{syllable}$)를 입력으로 사용하여 컨볼루션 신경망에 넣어주고 최대 풀링을 통해 하나의 고정된 크기의 벡터로 생성한다.

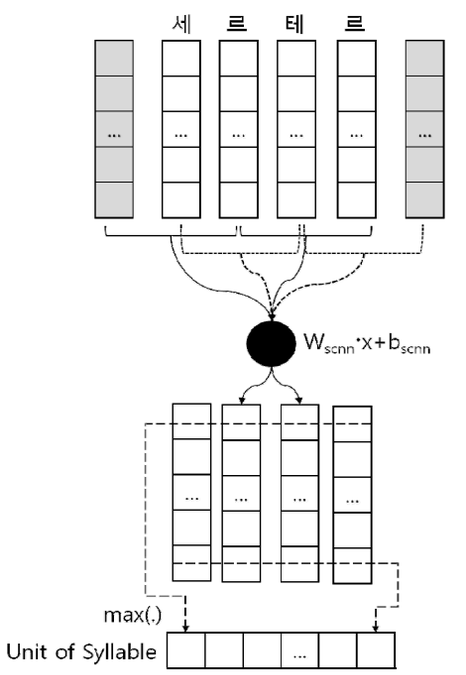

Fig. 3. Representation of a Word using Syllable-Unit based Vector

본 논문에서는 형태소 표현을 형태소의 단어 벡터와 품사 벡터, 그리고 음절 벡터를 합쳐서 표현하고 각각의 성능을 살펴보기 위해 Fig. 4와 같이 3가지로 나누어 비교를 수행한다. 첫째는 형태소 단어 벡터(e_{morph})와 품사 벡터(e_{pos})를 결합하여 형태소 벡터를 재구성한다(Fig. 4-a). 둘째는 형태소를 음절 벡터($e_{syllable}$)의 조합으로 표현하고 이를 품사 벡터(e_{pos})와 결합하여 형태소 벡터를 재구성한다(Fig. 4-b). 셋째, 형태소

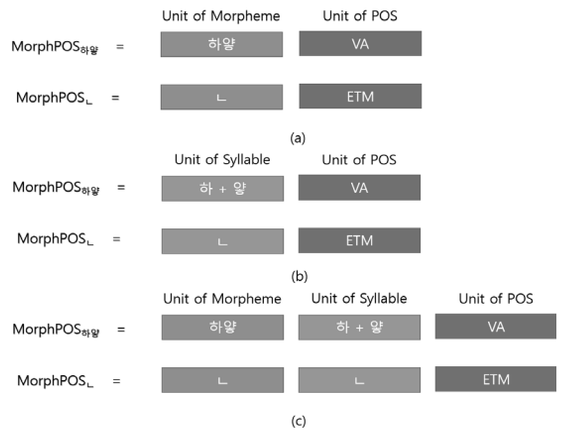

Fig. 4. The 3 Different Representations for MorphPOS Vector

의 단어 벡터(e_{morph})와 음절 벡터($e_{syllable}$)의 조합 그리고 품사 벡터(e_{pos})를 결합하여 형태소 벡터를 재구성한다(Fig. 4-c). 이와 같이 재구성한 형태소 벡터는 MorphPOS로 표시한다.

3.2 문맥(n-gram) 정보를 반영한 형태소 분석열 점수 모델

형태소 분석열 점수 모델은 3.1절에서 구성한 형태소 임베딩 벡터에 대한 문맥 정보를 활용한다. 이를 위해 컨볼루션 신경망을 이용하며 형태소 분석열의 길이 또한 다양하기 때문에 최대 풀링을 통해 하나의 고정된 크기의 벡터를 얻어낸다. 문맥 정보를 반영한 형태소 분석열 점수 모델의 수식은 Equation (6), (7), (8), (9)와 같다. Equation (6)은 컨볼루션 신경망의 입력으로 3.1절에서 제안한 형태소 임베딩 벡터(MorphPOS)가 사용된다.

$$x_{1:N} = MorphPOS_1 \oplus MorphPOS_2 \oplus \dots \oplus MorphPOS_N \quad (6)$$

Equation (7)은 인접한 형태소들에 대한 컨볼루션 계산이다. W는 필터의 파라미터이며, h는 필터의 범위로 n-gram의 n 값이다[17]. 본 연구에서는 n값을 5로 사용하였다. Equation (8)은 문장을 하나의 고정된 크기의 벡터로 표현하기 위한 최대 풀링 연산이다.

$$c_i = f(W \bullet x_{i:i+h-1} + b) \\ W \in R^{h \times d}; b \in R; f: non-linearity \quad (7)$$

$$Output_{CNN} = Max(c_1, c_2, \dots, c_{n-h+1}) \quad (8)$$

마지막으로 형태소 분석열에 대해 하나의 점수를 부여하기 위해 Equation (8)의 결과에 대한 선형 계층의 수식이다. (Equation 9)

$$Score = W_h \bullet Output_{CNN} + b_h \quad (9)$$

재순위화 모델의 학습을 위한 목적 함수는 최대-마진(Max-margin)을 사용한다[18]. 이 목적 함수는 정답 형태소 분석열의 점수인 $g(S_{answer})$가 정답이 아닌 형태소 분석열의 점수인 $g(S_{notanswer})$보다 높아지도록 파라미터들을 학습한다. 이때, 목적 함수의 수식은 Equation (10)과 같다. g의 함수는 재순위화 모델에서 나온 점수 값이다.

학습 파라미터 θ는 'e_{morph}, $e_{syllable}$, e_{pos}, W_{SCNN}, b_{SCNN}, W, b, W_h, b_h'이다.

$$J(\theta) = \max(0, 1 - g(S_{answer}) + g(S_{notanswer})) + \lambda \| \theta \|_2^2 \quad (10)$$

4. 실험 및 평가

4.1 실험 데이터

제안한 재순위화 모델은 기존 형태소 분석기의 후처리를 위한 모델로 Seq2seq 모델의 학습이 필요하다. 이를 위해 세종 코퍼스[1]의 문장 중 9만 개의 학습 문장과 9,997개의 문장을 평가 데이터로 사용하였다. 재순위화 모델의 학습을 위해 Seq2seq 모델에 학습 데이터를 입력으로 넣어주고 10개의 후보 형태소 분석열을 생성한다. 이 중 정답이 존재하는 경우만 추출하여 79,418개의 문장을 재순위화 모델의 학습 데이터로 사용하였다. 학습 데이터는 (정답 형태소 분석열, 오답 형태소 분석열) 쌍으로 구성하며, 이를 이용하여 3장에서 제안한 모델을 학습시켰다. 평가 데이터는 Seq2seq 모델에서 사용한 평가 데이터를 그대로 사용한다. 학습 데이터와 평가 데이터에 대한 정보를 Table 1에 제시하였다.

Table 1. Description of Train and Test Data [4]

Model	Train		
	number of sentences	Avg. number of words	Avg. number of morphemes
Seq2seq	90,000	14.56	47.33
Reranking	79,418	12.94	41.76
Model	Test		
	number of sentences	Avg. number of words	Avg. number of morphemes
Seq2seq & Reranking	9,997	13.88	45.71

4.2 모델 학습을 위한 파라미터

Table 2은 Seq2seq 모델을 학습하기 위한 파라미터 정보이다. 파라미터의 값은 [3]의 연구 모델에서 제시한 파라미터 정보를 사용하였다. Seq2seq 모델의 학습은 pytorch 버전의 OpenNMT[2] 라이브러리를 활용하였다.

Table 3는 재순위화 모델을 학습하기 위한 파라미터 정보이다.

1) https://ithub.korean.go.kr/user/main.do
2) https://github.com/OpenNMT/OpenNMT-py

Table 2. Hyperparameters for Training Seq2seq Model

Parameter	Value
Embedding dimension	256
RNN Type	GRU
RNN layers	2
RNN size	256
Bidirectional	True
Bidirectional Merge	concat
Optimizer	Adam
Learning rate	1e-3

Table 3. Hyperparameters of Reranking Model

Parameter	Value
Morph Embedding Dimension	100
Syllable Embedding Dimension	256
POS Embedding Dimension	25
Syllable Convolution Kernel	3
n-gram Convolution Kernel	5
Convolution Activation function	ReLU
Numbers of Linear Layer	2
Linear Activation function	tanh
Optimizer	Adam
Learning rate	1e-3
Weight decay	1e-5

4.3 실험 결과

성능 평가는 형태소 단위의 F1 점수를 사용한다. 3.1절에서 형태소의 표현을 3가지 방법으로 제안하였다. Table 4는 3.1절에서 제안한 3가지 방법에 대한 실험 결과이다. 형태소 단어와 형태소의 음절 그리고 품사를 결합하여 입력으로 사용하였을 때 96.21%의 성능이 나왔으며, 가장 높은 실험결과를 보여주었다.

Table 4. The Experimental Result of Embedding Representations

Embedding Representations	Reranking Model
Morphemes + POS	96.00%
Syllables + POS	95.59%
Morphemes + Syllables + POS	**96.21%**

Table 5는 Seq2seq 모델의 형태소 분석기(Baseline)와 후처리 방법으로 재순위화 모델(Reranking)을 통해 나온 형태소 분석기의 실험 결과이다. 후처리 방법으로 제안한 재순위화 모델은 기존 형태소 분석기에 비해 1.17%의 F1 점수가 향상되었다.

Table 5. The Performance Result of Reranking Model

Model	F1-score
Baseline(Seq2seq)[3]	95.04%
Reranking	**96.21%**

(예 3)은 재순위화 모델을 통해 Seq2seq 모델에서 가지고 있던 오류를 해결한 예제들이다.

(1)은 '넘나'라는 대명사는 사전에 존재하지 않는다. 재순위화 모델을 통해 '넘나들다'라는 동사 형태로 오류를 해결한 것을 볼 수 있다. (2)는 '지니다'로 원형 복원을 못한 오류를 재순위화 모델을 통해 해결한 것을 볼 수 있다. (3)은 '만보'라는 사전에 존재하지 않는 동사를 생성했는데 이는 Seq2seq 모델의 입력과 출력이 음절 단위로 나오기 때문에 생긴 문제이다. 형태소 단위의 정보를 활용한 재순위화 모델에서 오류를 해결한 것을 확인할 수 있었다. (4)는 '글씨'라는 형태소에서 '씨'를 의존 명사로 보는 것과 '글씨' 자체를 명사로 보는 것의 중의성 문제를 포함하고 있다. 이 또한 형태소 단위의 문맥 정보를 활용하여 오류를 해결한 것을 볼 수 있다.

(예 3)

(1) (seq2seq) 일상적 실존을 넘나/NP+들/XSN 가능성이 없다면 ⇒ (reranking) 넘나들/VV+ㄹ/ETM

(2) (seq2seq) 호소력을 지녀/VV+어야/EC 한다 ⇒ (reranking) 지니/VV+어야/EC

(3) (seq2seq) 허바드 대사와 만보/VV+았/EP+다/EF ⇒ (reranking) 만나/VV+았/EP+다/EF

(4) (seq2seq) 좌수로 글/NNG+씨/NNB+를/JKO 쓸 수밖에 ⇒ (reranking) 글씨/NNG+를/JKO

반면에 Seq2seq 모델에서는 정답으로 출력했지만 재순위화 모델에 의해 오답 결과를 1순위로 출력한 경우도 발생하였다. (예 4)는 재순위화 과정에서 발생한 오류에 대한 예제이다. 이 오류들은 주로 재순위화 모델에서 원문의 정보를 참조하지 못하기 때문에 생긴 문제이다. 재순위화 모델에서는 K 개의 후보 결과 중, 형태소 정보와 형태소 사이의 문맥 정보만을 갖고 순위를 조정하다보니 원문과 다른 결과가 높은 순위를 갖게 되는 경우가 발생하였다.

(예 4)

(1) (seq2seq) 함성을 지르/VV+었/EP+다/EF ⇒ (reranking) 누르/VV+었/EP+다/EF

(2) (seq2seq) 말문을 열/VV+ㄴ/ETM 정 차관은 ⇒ (reranking) 타/VV+ㄴ/ETM

5. 결 론

본 연구에서는 기존 Seq2seq 모델 기반의 한국어 형태소 분석기의 결과에 대해 재순위화 모델을 사용하여 형태소 분

석 결과의 순위를 재조정하는 후처리 방법을 제안하였다. 형태소 분석 결과의 순위를 재조정하기 위해 Seq2seq 모델의 디코더 단계에서 빔 서치 기법을 사용하여 여러 개의 후보 형태소 분석 결과를 생성하였다. 각각의 후보 형태소 분석 결과에 대해 형태소 단위의 문맥 정보를 활용하여 형태소 분석 결과의 순위를 재조정하였다.

음절 단위의 형태소 분석 결과를 형태소 단위의 표현으로 재구성하기 위해 세 가지의 방법을 제안하였다. 첫째는 형태소 단어 임베딩과 품사 임베딩의 결합이고, 둘째는 형태소의 음절 임베딩과 품사 임베딩의 결합이다. 셋째는 형태소의 단어 임베딩, 형태소의 음절 임베딩 그리고 품사 임베딩의 결합이다. 형태소의 음절 임베딩의 경우, 고정된 크기의 벡터로 표현하기 위해 컨볼루션 신경망을 사용하였다.

형태소 단위로 표현된 분석 결과를 형태소 단위의 문맥 정보를 활용하여 점수로 환산할 수 있는 신경망을 구축하고 학습시켰다. 실험 결과 형태소를 단어, 음절, 품사의 결합 벡터로 표현했을 때, 96.21%의 성능을 보였으며 이는 기존의 Seq2seq 모델 한국어 형태소 분석기에 비해 1.17%의 F1 점수가 향상된 것이다. 그렇지만 Seq2seq 모델이 출력한 정답 형태소 분석열을 오답으로 출력한 경우도 있었다. 이는 재순위화 모델에서는 원문의 정보를 반영하지 못하다 보니 생기는 오류였다. 이 문제는 향후 연구를 통해 해결해 보고자 한다.

References

[1] S. H. Na and S. K. Jung, “Deep Learning for Korean POS Tagging,” *Korea Software Congress*, pp.426-428, 2014.

[2] J. LI, E. H. Lee, and J. H. Lee, “Sequence-to-sequence based Morphological Analysis and Part-Of-Speech Tagging for Korean Language with Convolutional Features,” KIISE, Vol.44, No.1, pp.57-62. 2017.

[3] H. S. Hwang and C. K. Lee, “Korean Morphological Analysis using Sequence-to-sequence learning with Copying mechanism,” *Korea Software Congress*, pp.443-445, 2016.

[4] Y. S. Choi, “Re-ranking Model of Korean Morphological Analysis Results using Neural Networks,” Master's thesis, Chungnam National University, Republic of Korea, 2017.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” *Advances in neural information processing systems*, pp.3104-3112, 2014.

[6] J. Gu, Z. Lu, H. Li, and V. O. Li, “Incorporating copying mechanism in sequence-to-sequence learning,” arXiv preprint arXiv:1603.06393, 2016.

[7] F. J. Och and H. Ney, “The alignment template approach to statistical machine translation,” *Computational Linguistics*, Vol.30, No.4, pp.417-449, 2004.

[8] J. Li and D. Jurafsky, “Mutual information and diverse decoding improve neural machine translation,” arXiv preprint arXiv:1601.00372, 2016.

[9] O. Dušek and F. Jurčíček. “Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings,” arXiv preprint arXiv:1606.05491, 2016.

[10] J. Chorowski and J. Navdeep, “Towards better decoding and language model integration in sequence to sequence models,” arXiv preprint arXiv:1612.02695, 2016.

[11] Y. Shao, S. Gouws, D. Britz, A. Goldie, B. Strope and R. Kurzwell, “Generating high-quality and informative conversation responses with sequence-to-sequence models,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp.2200-2209, 2017.

[12] T. H. Wen, M. Gasic, D. Kim, N. Mrksic, P. H. Su, D. Vandyke, and S. Young, “Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking,” arXiv preprint arXiv:1508.01755, 2015.

[13] T. Mikolov, W. T. Yih, and G. Zweig, “Linguistic Regularities in Continuous Space Word Representations,” in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.746-751, 2013.

[14] C. D. Santos and B. Zadrozny, “Learning character-level representations for part-of-speech tagging,” in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp.1818-1826, 2014.

[15] S. Misawa, M. Taniguchi, and Y. Miura, “Character-based Bidirectional LSTM-CRF with words and characters for Japanese Named Entity Recognition,” in *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pp.97-102, 2017.

[16] T. Nakagawa and K. Uchimoto, “A hybrid approach to word segmentation and pos tagging,” in Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. Association for Computational Linguistics, pp.217-220, 2007.

[17] Y. Kim, “Convolutional neural networks for sentence classification,” arXiv preprint arXiv:1408.5882, 2014.

[18] L. Rosasco, E. D. Vito, A. Caponnetto, M. Piana, and A. Verri, “Are loss functions all the same?” *Neural Computation*, Vol.16, No.5, pp.1063-1076, 2004.

최 용 석

https://orcid.org/0000-0002-7889-8004

e-mail : yongseok.choi.92@gmail.com

2016년 충남대학교 정보통신공학과(학사)

2018년 충남대학교 전자전파정보통신공학과(석사)

2018년~현 재 충남대학교 전자전파정보통신공학과 박사과정

관심분야 : 자연언어처리, 정보검색, 기계학습, 인공지능

이 공 주

https://orcid.org/0000-0003-0025-4230

e-mail : kjoolee@cnu.ac.kr

1992년 서강대학교 전자계산학과(학사)

1994년 한국과학기술원 전산학과(공학석사)

1998년 한국과학기술원 전산학과(공학박사)

1998년~2003년 한국마이크로소프트(유) 연구원

2003년 이화여자대학교 컴퓨터학과 대우전임강사

2004년 경인여자대학 전산정보과 전임강사

2005년~현 재 충남대학교 전파정보통신공학과 교수

관심분야 : 자연언어처리, 기계번역, 정보검색, 정보추출