음소 단위 임베딩 기반 한국어 모델

최우성*, 현경석*, 정재화**, 정순영*†
*고려대학교 컴퓨터학과
**한국방송통신대학교 컴퓨터과학과
e-mail: ws_choi@korea.ac.kr, ks_hyun@korea.ac.kr, jaehwachung@knou.ac.kr, jsy@korea.ac.kr

Phoneme-level Embedding based Korean Language Model

Woosung Choi*, Kyungseok Hyun*, Jaehwa Chung**, Soon Young Jung*+
*Dept of Computer Science, Korea University
**Dept of Computer Science, Korea National Open University

요 약

최근 제안되고 있는 Bert 등의 딥러닝 언어 모델 기반 pre-training 기법은 다양한 NLP 분야에서 활용되고 있다. 텍스트로 작성된 데이터 셋을 딥러닝 언어 모델이 학습하기 위해서는 토크나이징(tokenizing) 기술이 필요하다. 그러나 기존 토크나이징 방식은 한국어 및 한글이 가지는 고유한 특성(교착어적 특성과 모아쓰기 반영)을 반영하기 어렵다는 한계를 가지고 있다. 본 논문에서는 한국어와 한글이 가지는 고유한 특성을 고려하기 위하여 음소 단위의 임베딩 기법을 제안하며, 이를 기반으로 언어 모델을 설계 및 구현한다. 또한 음소 단위 임베딩 기반 한국어 모델이 실제 데이터 집합(구약성서)에서 나타나는 언어적 패턴을 학습할 수 있다는 것을 실험을 통하여 밝힌다.

1. 서론

최근 제안되고 있는 Bert[1] 등의 신경망 기반 언어 모델은 다양한 NLP(Natural Language Processing) 분야에서 활용되고 있다. 신경망 기반 언어 모델을 대규모 텍스트 데이터 집합을 이용해 학습시키면, 대상 언어의 특성을 반영한 표상 (learned representation)을 얻을 수 있다. 이러한 표상을 이용하여 특정 NLP 과제(task)를 해결하는 모델을 만들 경우, One-hot-encoding과 같은 단순한 표상을 이용할 때보다 더 높은 정확도를 얻을 수 있다[2].

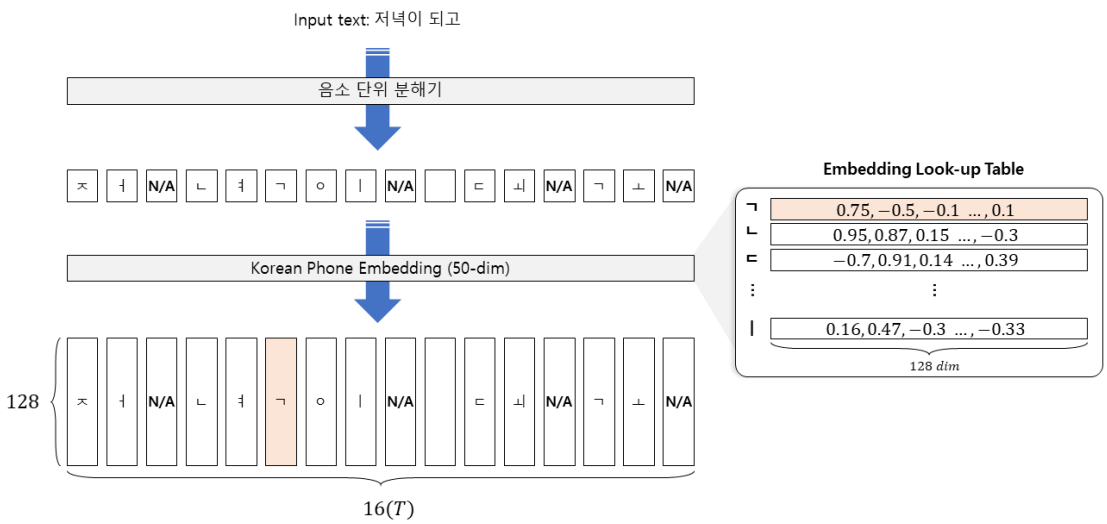
언어 모델의 종류에는 주어진 n개의 연속된 단어를 분석하여 다음 순서의 단어를 예측하는 순차적 모델[3], 주변 C개의 단어를 분석하여 가운데 단어를 예측하는 모델[2], 임의의 단어를 가린(Masking)한 후, 다른 단어들을 이용하여 해당 단어를 예측하는 Masked 언어 모델[1] 등이 있다. 최근에는 더 효율적이고 유연한 언어처리를 위해 단어보다 더 작은 단위를 입출력으로 사용하는 방법이 제안되고 있다. 이를테면 데이터 기반으로 학습된 sub-word를 이용하는 방법(wordpiece[4], sentencepiece[5])이 제안되었으며, 문자 자체를 대상으로 하는 Character-aware 언어 모델[6] 또한 제안되었다.

특히, 문자 단위로 문장을 받아들이는 언어 모델을 제안한 [6]의 경우 영어보다 형태소(morphology)가 풍부한 언어(Arabiac, Czech, French, German, Spanish, Russian)에 대해 더욱 유연하게 대처할 수 있음을 실험적으로 밝혔다. 실험결과, 문자 단위로 입력텍스트를 처리하는 [6]의 언어 모델은 단어 또는 형태소 단위로 입력을 처리하는 언어 모델에 비해 형태소가 많은 언어를 대상으로 한 훈련결과 더 낮은 perplexity 값을 보였다.

이러한 실험결과를 토대로 보았을 때, 형태소가 복잡한 한국어[7] 모델 설계 시 단어보다 더 작은 단위의 토큰을 사용하는 것이 더 바람직하다. 예를 들어 형태소 분석을 통해 한국어 문장을 분해하고, 분해된 형태소 단위로 언어 모델을 설계하는 방식, 한국어 어절 단위로 설계하는 방식 등[8]이 있다.

그러나 형태소 분석 전처리를 이용한 토큰화 방법이나, 어절 단위의 언어 모델 방식은 한국어 및 한글이 가지는 특성을 반영하지 못한다는 한계점을 가지고 있다. 전자의 경우 형태소 분석기의 성능에 의존적인 구조로 되어 있다. 또한, 형태소 분석기는 일반적으로 지도학습(supervised learning)을 기반으로 구축되기 때문에, 언어 모델 기반의 비지도 학습(supervised learning) 과 병행한 end-to-end 학습이 어렵다는 문제가 있다. 후자의 경우 모아쓰기 방식을 사용하여 어절 단위로 문자를 표현하는 한글의 특성으로 인해 한국어 고유의 언어적 특성을 반영할 수 없다는 문제가 있다. 이러한 문제를 해결하기 위해 본 논문에서는 음소단위로 한국어를 모델링하는 신경망 기반 기법을 제안한다.

† 교신저자: 정순영 고려대학교 jsy@korea.ac.kr
이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2016R1D1A1B03930907). 또한 이 논문은 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2019R1F1A1062719)



(그림 1) 음소단위 임베딩 기법

2. 관련연구

이러한 문제를 해결하기 위해 본 논문에서는 음소단위로 한국어를 모델링하는 신경망 기반 기법을 제안한다. 비슷한 기존 기법인 [10]에서는 one-hot-encoding 기반의 음소 단위 언어 모델링을 수행하였으나, 본 연구에서는 learned embedding 방식을 기반으로 음소 단위 언어 모델링을 구현한다.

3. 음소 단위 임베딩 기반 한국어 모델

본 논문에서는 [9]와 유사한 음소 단위 한국어 모델을 제안한다. 본 연구에서 제안하는 한국어 모델에서는 음소 단위로 텍스트를 분해한 후 연속된 음소 T개를 읽고 그 뒤에 나타날 음소를 예측한다다. 언어 모델에 대한 형식적 정의는 다음과 같다: 본 논문에서는 음소 시퀀스 $phs = (ph_1, ph_2, \ldots, ph_T)$에 대해

$$\hat{P}(phs) = \hat{P}(ph_T | ph_1 \ldots ph_{T-1})$$

를 학습하는 신경망 기반 언어모델을 제안한다. 이러한 모델 구현을 위해 본 논문에서는 음소 단위 임베딩을 이용하여 각 음소를 벡터화하며, 벡터화된 음소열(sequence)를 RNN(Recurrent Neural Network)에 순서대로 입력함으로써 $(ph_1, ph_2 \ldots, p_{T-1})$에 대한 표상(representation)을 얻는다. 이렇게 얻은 표상을 기반으로, 각 음소에 대해 ph_T일 확률을 구한다. 각각에 대한 자세한 설명은 다음과 같다.

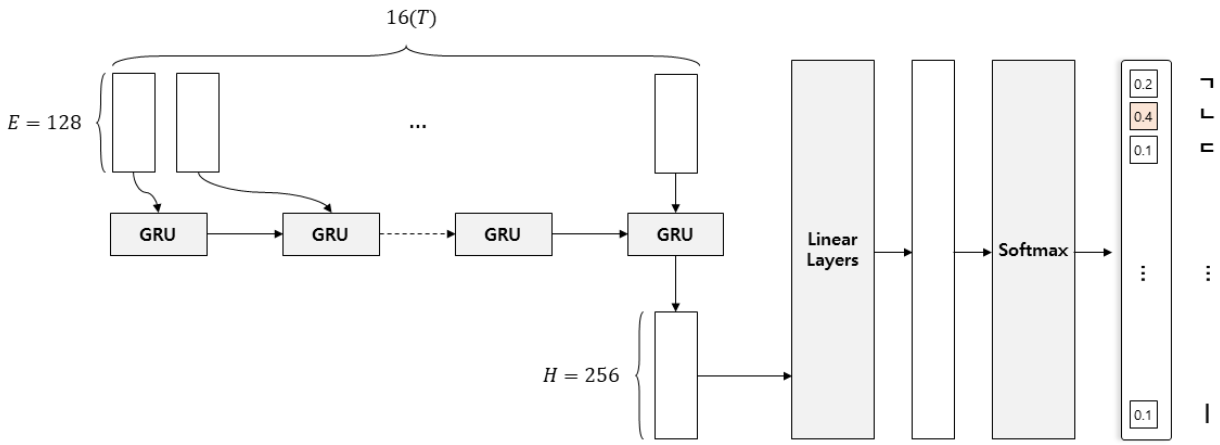
3.1. 음소 단위 임베딩

본 논문에서 구현한 음소단위 임베딩 구조는 다음과 같다. (그림 1)과 같이 '저녁이 되고'라는 텍스트가 주어진 경우, 해당 텍스트를 음소단위로 분해한다. 이때, 분해는 한국어 모아쓰기 음절 단위로 이루어지며, 각 음절은 3개의 음소로 분해된다. '녁'의 경우 초성인 'ㄴ', 중성인 'ㅕ', 종성인 'ㄱ'으로 구분된다. '저'의 경우에는 초성인 'ㅈ', 중성인 'ㅓ', 종성의 경우 존재하지 않는다는 뜻에서 'N/A'으로 표기하였다. 이와 같은 방법으로 '저녁이 되고'는 'ㅈ','ㅓ','N/A',...,'ㄱ','ㅗ','N/A'와 같은 16개의 음소열로 분해된다.

분해된 음소열은 Korean Phone Embedding Layer를 통해 E차원으로 embedding된다. 각 음소열에 대해 E차원 (e.g. 128차원)의 임의 벡터값을 할당하였다. (그림 1)의 오른쪽 행렬은 공백문자와 'N/A'를 포함한 존재할 수 있는 모든 한글 자모음 53개에 대해 할당된 E(=128) 차원의 벡터를 쌓아 만든 look-up 행렬이다. 음소 단위 분해기를 통해 16개의 음소열로 바뀐 입력텍스트는, 더사 Korean phone Embedding Layer를 거쳐 각각이 E차원인 16개의 벡터열로 변형된다. 이렇게 변형된 벡터는 3.2에서 다룰 RNN 기반의 신경망 기반 시계열 모델에 입력되어 전체 문장에 대한 표상을 얻을 수 있다.

3.2. Recurrent Neural Network 기반 벡터화

본 논문에서는 RNN의 변형 중 장거리 기억력 보존성이 뛰어난 GRU(Gated Recurrent Unit)[9]을 사용하여 주어진 음소열을 하나의 벡터값으로 수치화한다. GRU는 RNN과 마찬가지로, n-1번째의 은닉 출력이 n번째 입력값과 함께 합산되어 n번재 은닉 출력을 생성하는 순환(recurrent) 구조로 되어 있으므로, 순차적인 데이터를 모델링하기 적합하다. (그림 2)와 같이, 3.1의 결과물인 벡터열은 차례대로 GRU 모델에 순서대로 입력된다. 본 논문에서는 가장 마지막 음소열을 입력한 후의 GRU 출력값((그림 2)의 H 차원 벡터)을 해당 음소열을 대표하는 표상(representation)으로 삼았다.



(그림 2) 음소단위 임베딩 기법

이렇게 얻은 표상은 전체 음소열, 즉 $(ph_1, ph_2, \ldots ph_{T-1})$의 특성을 모두 반영한 표상이다. 본 논문은 이 표상을 바탕으로 ph_T를 예측하는 모델을 제안한다. ph_T를 예측하는 모델은 3.3절에서 자세히 다룬다.

3.3. 음소 예측 모델

음소예측 모델의 경우, 음소열 $(ph_1, ph_2, \ldots ph_{T-1})$를 대표하는 H차원 벡터를 입력받아 각 53개의 음소별 출연 확률을 계산한다. 이를 위해 H차원을 53차원으로 변환하는 신경망 구조를 설계하였다. (그림 2)의 오른쪽 부분은 본 논문에서 구현한 음소예측 모델이다. H차원의 GRU 출력인 GRU_{output}를 입력받았을 때의 출력은 다음과 같은 형식으로 정의된다.

$$ps = softmax(W * GRU_{output} + b)$$

(단, $W \in R^{53 \times H}$이며 $b \in R^{53}$)

이렇게 정의된 출력 ps의 i번째 차원, 즉, ps_i는 i번째 음소가 $(ph_1, ph_2, \ldots ph_{T-1})$ 뒤에 나타날 확률을 뜻한다.

4. 실험 및 분석

본 논문에서 제안한 언어 모델이 실제 데이터 집합의 특성을 잘 파악할 수 있는지 확인하기 위해 [10]에서와 같이 구약성경 문장생성 실험(3,1)을 수행하였다. 또한 각 음소가 어떻게 임베딩되었는지 가시화(3.2)하였다.

공백문자와 'N/A'를 포함한 존재할 수 있는 모든 한글 자모음 53개 및 특수문자 117개를 고유 토큰으로 삼았으며 연속된 (T-1)개의 음소를 읽고 다음에 위치할 음소를 예측하였다. T로는 매 epoch마다 50에서 200 사이의 임의의 값을 배정하였다. 각 음소에 대해 다음 음소 발생 확률을 계산하고, Negative Log-Likellyhood 값을 최소화하는 방향으로, 수렴할 때까지 모델을 학습시켰다.

실험에 사용된 장비는 다음과 같다.

<표 1> 사용 장비

구분	품목
CPU	Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz
GPU	NVIDIA GeForce RTX 2080Ti
RAM	DDR4 4G x 4

본 논문에서는 PyTorch를 이용하여 제한 언어 모델 구현했으며, 사용된 하이퍼 파라미터는 다음과 같다.

<표 2> Hyper parameter 목록

기호	의미	예시
E	음소 임베딩 차원 수	128
H	GRU 은닉차원 수	256
L	GRU 계층 수	4
t	temperature(문장생성)	0.5

E, H, L의 경우 (그림 1)과 (그림 2)에서 명시된 모델 구축을 위한 hyper parameter다. t의 경우 문장생성을 위해 사용되는 파라미터로, t가 낮을수록 생성된 텍스트가 학습 데이터의 일부분과 완전히 일치할 확률이 높아지며, t가 높을수록 예측의 의외성이 강화되어 원본 데이터와 완전히 일치하는 문장을 찾을 확률이 낮아진다. 각각의 실험에 대한 결과는 다음과 같다.

4.1. 문장 생성 실험

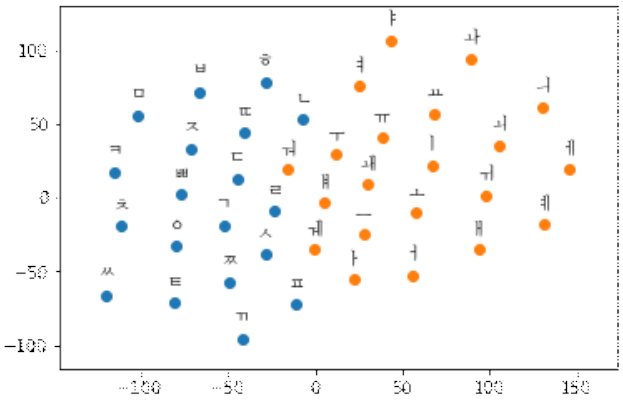
구약성서로 학습한 제안 언어 모델에 '여호'라는 입력을 주고, 이를 기반으로 텍스트를 생성하게 하였다. 결과는 <표 3>과 같다. 실험결과 비교적 적은 매개변수를 가지는 모델(<표 1>의 (32,64,2) 모델)로도 구약성서와 유사한 텍스트를 생성해낼 수 있음을 볼 수 있다.

<표 3> 문장생성 결과

모델 (E,H,L)	t	생성문장
(128,256,4)	0.4	여호와께서 나를 위하여 그 아래에서 그 다음을 듣고 성읍에 가서 나오미의 영광을 받은 자들을 위하여 그 앞에 있는 모든 일을 행하여 이르되 여호와의 말씀이니라
	0.8	여호와가 너희를 나담하셨으니 나는 모든 사람은 아는 보는 것이요 여호와의 항하는 유다 속에 들이지도 아니하니라 그러나 네가 앗수르 아들이라 하였더니 내가 탐사지 말라 하거늘
(32,64,2)	0.4	여호와께서 내가 이르되 여호와의 아들 그들이 나의 아를 따라 이르되 그러라 하시니라
	0.8	여호와의 그들이 그를 기탐을 그 주의 죽인 것을 내게 보은 일 아니하였느니라

4.2. 임베딩 결과 분석 및 가시화

구약성서로 학습한 제안 언어 모델은 각 음소에 할당된 임베딩 벡터 또한 동시에 학습한다. (그림 3)은 수렴한 (128, 256, 4) 모델의 각 음소에 대한 임베딩을 t-SNE[11]를 이용하여 2차원으로 가시화한 그림이다.



(그림 3) 음소 임베딩 가시화

그림에서 볼 수 있듯, 자음은 주로 제 2,, 제 3사분면, 모음은 주로 제 1, 제 4사분면에 분포한 것을 확인할 수 있다. 임베딩이 자음과 모음의 특성을 구분지을 수 있는 것으로 보아, 제안 모델의 음소 임베딩은 음소에 대한 사전 지식을 반영하지 않고 설계되었음에도, 이를 데이터로부터 학습하였다고 볼 수 있다.

5. 결론 및 향후 연구

본 논문에서는 음소 단위 임베딩 기반 한국어 모델을 설계 및 구현하였으며, 구약성서 텍스트 생성 실험을 통해 제안 모델이 입력된 데이터 집합의 언어적 패턴을 학습할 수 있다는 것을 밝혔다. 또한 각 음소에 대한 임베딩 벡터를 가시화해봄으로써 제안 모델이 각 음소의 특징을 데이터로부터 할 수 있음을 보였다. 향후 연구 주제로는 대요량 텍스트 데이터에 대한 사전학습(pre-training)을 통해 얻은 임베딩을 이용하여 End-to-End NLP 시스템을 연구 및 개발할 예정이다.

참고문헌

[1] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[2] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

[3] Bengio, Yoshua, et al. "A neural probabilistic language model." Journal of machine learning research 3.Feb (2003): 1137-1155.

[4] Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." arXiv preprint arXiv:1609.08144 (2016).

[5] Kudo, Taku, and John Richardson. "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing." arXiv preprint arXiv:1808.06226 (2018).

[6] Kim, Yoon, et al. "Character-aware neural language models." Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[7] 하김병창, 이근배, and 이종혁. "한국어 연속음성인식을 위한 Viterbi 형태소 분석기." 정보과학회논문지 (B) 24.12 (1997): 1493-1501.

[8] 정상근. "심층신경망에서의 효과적인 문장추상화를 위한 한국어 단위 연구." 정보과학회논문지 45.5 (2018): 457-465.

[9]Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

[10] 안성만, et al. "한국어 음소 단위 LSTM 언어모델을 이용한 문장 생성." 지능정보연구 23.2 (2017): 71-88.

[11] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.Nov (2008): 2579-2605.