The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.10, pp.553~558
eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.10.553

Syllable-Level Lightweight Korean POS Tagger using Transformer Encoder

Suyoung Min† · Youngjoong Ko††

ABSTRACT

Morphological analysis involves segmenting morphemes, the smallest units of meaning or grammatical function in a language, and assigning part-of-speech tags to each morpheme. It plays a critical role in various natural language processing tasks, such as named entity recognition and dependency parsing. Much of modern natural language processing relies on deep learning-based language models, and Korean morphological analysis can be broadly categorized into sequence-to-sequence methods and sequential labeling methods. This study proposes a morphological analysis approach using the transformer encoder for sequential labeling to perform syllable-level part-of-speech tagging, followed by morpheme restoration and tagging through a pre-analyzed dictionary. Additionally, the CBOW method was used to extract syllable-level embeddings in lower dimensions, designing a lightweight morphological analyzer model with reduced parameters. The proposed model achieves fast inference speed and low parameter usage, making it efficient for use in resource-constrained environments.

Keywords : Transformer Encoder, Morphological Analysis, Part of Speech Tagging, Sequential-Labeling

트랜스포머 인코더를 활용한 음절 단위 경량화 형태소 분석기

민 수 영† · 고 영 중††

요 약

형태소 분석은 의미를 가지거나 문법적 기능을 하는 언어의 최소 단위인 형태소를 분리하고, 각 형태소의 품사를 결정하는 작업으로, 개체명 인식, 의존구문 분석 등과 같은 자연어 처리 작업에서 중요한 역할을 한다. 현대 자연어처리의 많은 부분은 딥러닝 기반 언어 모델에 의존하고 있으며, 딥러닝 기반 한국어 형태소 분석은 크게 시퀀스-투-시퀀스 방식과 순차적 레이블링 방식으로 나뉜다. 본 연구에서는 트랜스포머 인코더를 활용한 순차적 레이블링 방식으로 음절 단위 품사 태깅을 수행한 후, 기분석 사전을 통해 형태소 복원 및 품사 태깅을 진행하는 형태소 분석 방식을 제안한다. 또한, CBOW 방식을 사용하여 음절 단위 임베딩을 낮은 차원으로 추출함으로써 파라미터 수를 줄인 경량화 형태소 분석기 모델을 설계하였다. 제안된 모델은 낮은 파라미터 수와 빠른 추론 속도를 통해 자원이 제한된 환경에서도 효율적으로 활용될 수 있다.

키워드 : 트랜스포머 인코더, 형태소 분석기, 형태소 품사 태깅, 순차적 레이블링

1. 서 론

형태소 분석은 컴퓨터 언어 처리 분야에서 주요한 연구 주제 중 하나로, 자연어의 기본 의미 단위인 형태소를 식별하고 적절한 품사를 할당하는 과정이다. 형태소 분석은 다른 언어 분석에서 전처리된 입력으로 활용되기 때문에, 빠른 처리 속도와 높은 정확성이 요구된다. 그러나 한국어 형태소 분석은 영어와 같은 언어 분석과 달리 교착어인 특성으로 인해 몇 가지 결점들이 존재하는데, 그중 하나는 한국어에 있는 여러 불규칙 활용으로 입력 문장과 형태소 분석 결과의 형태와 길이가 일치하지 않는 점이다. 이러한 입력과 출력 길이의 불일치 문제가 있기에 필연적으로 정확도와 효율성 저하가 발생하게 된다.

딥러닝 기반의 한국어 형태소 분석 관련 연구들은 크게 두 가지 방식이 있다. 첫 번째로는 문장 각 음절에 후보 품사 태그를 부착하여 이후 형태소 단위로 복원 및 형태소에 적절한 품사를 부착하는 순차적 레이블링(Sequence Labeling) 방식[1-4], 두 번째로는 인코더-디코더(Encoder-Decoder) 구조를 활용한 시퀀스-투-시퀀스(Sequence-to-Sequence) 방식[5-10]으로, 인코더가 전달하는 문맥 관련 정보를 활용하여 디코더

※ 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(No. RS-2024-00350379, 40)와 과학기술정보통신부 및 정보통신기획평가원의 생성AI선도인재양성사업(IITP-2024-RS-2024-00360227, 40), 그리고 정부(과학기술정보통신부)의 재원으로 정보통신 기획평가원의 지원을 받아 수행된 ICT명품인재양성사업(RS-2020-II201821, 20)의 연구 결과로 수행되었음.
† 준 회 원 : 성균관대학교 AI시스템공학과 석사과정
†† 비 회 원 : 성균관대학교 소프트웨어학과 교수
Manuscript Received : August 16, 2024
Accepted : September 2, 2024
* Corresponding Author : Youngjoong Ko(youngjoong.ko@gmail.com)

가 형태소와 품사 쌍을 생성하는 방식이 있다.

본 연구에서는 주의 집중 기법(Attention Mechanism)[11]을 활용한 모델인 트랜스포머(Transformer) 인코더를 음절 단위 품사 태깅에 적용 및 낮은 차원의 임베딩을 활용하여 거대 언어 모델과 병행 사용이 가능한 순차적 레이블링 구조의 경량화 형태소 분석기 모델을 제안한다.

본 연구의 형태소 분석기는 두 가지 단계로 구성되어 있다. 첫 단계는 트랜스포머 인코더를 활용하여 음절에 후보 품사 태그를 부착하는 단계이고, 두 번째 단계는 기분석 사전 및 불규칙 변환 사전을 활용하여 후보 품사 태그가 부착된 음절들을 형태소 단위로 복원하고 품사를 부착하는 단계이다. 음절 단위 품사 태깅 단계에서는 공개된 신문 기사 데이터를 활용해 음절 단위의 임베딩을 추출하고, 이를 기반으로 트랜스포머 인코더를 학습시켜 각 음절에 순차적으로 후보 품사 태그를 할당한다. 그 후, 기분석 사전과 불규칙 변환 사전을 이용해 음절과 해당 품사 태그 쌍을 형태소 및 품사 쌍으로 복원하여 형태소 분석을 마무리한다.

현재 자연어처리의 추세인 거대 언어 모델들은 높은 단위의 파라미터 수와 대용량의 데이터로 훈련되어 높은 연산 자원과 메모리 용량을 필요로 한다. 따라서 일반적인 딥러닝 기반의 언어 분석기로는 거대 언어 모델과 병행하여 사용하기 힘들 수 있다. 본 연구의 형태소 분석기는 이러한 문제를 해결할 수 있는 빠른 추론 속도를 가진 경량화된 형태소 분석 모델 개발을 목적으로 한다.

2. 관련 연구

과거 자연어처리 분야에서는 순환신경망(Recurrent Neural Network; RNN)이나, 장단기 메모리(Long Short Term Memory; LSTM) 기반 모델들을 활용하여 시계열 데이터를 다루었다. 그러나 이런 모델들은 재귀적인 특성으로 문장의 길이에 따라서 성능 및 처리 속도에 큰 단점을 지니고 있다.

이와 같은 문제를 해결하기 위해 주의 집중 기법을 도입한 트랜스포머 모델[11]이 소개되었다. 트랜스포머 모델은 재귀적인 특성을 주의 집중 기법으로 대체하여 구성하였기 때문에 입력 데이터를 병렬적으로 처리할 수 있어 학습 과정을 빠르게 진행할 수 있다. 또한, 반드시 이전 단계의 정보가 필요한 재귀적 모델과는 다르게, 데이터의 위치에 대한 정보를 위치 인코딩(Positional Encoding)을 통해 기본적으로 순서가 없는 구조인 트랜스포머 모델에 위치 관련 정보를 추가하여 결점을 보완한다.

이렇게 입력된 시계열 데이터를 자기 주의 집중(Self Attention) 단계를 통해 입력 데이터 내의 각 단어끼리의 관련성에 대한 모델링을 Equation (1)과 같이 진행한다. 그러나 자기 주의 집중 계산 하나로는 관련성에 대한 정보가 부족할 수 있으니, 여러 관점에서 정보를 수집하는 다중 주의 집중(Multi-head Attention) 단계를 통해 다양한 관점에서 정보를 처리하여 데이터의 여러 특징을 포착하고, 풍부한 표현력을 달성할 수 있게 된다.

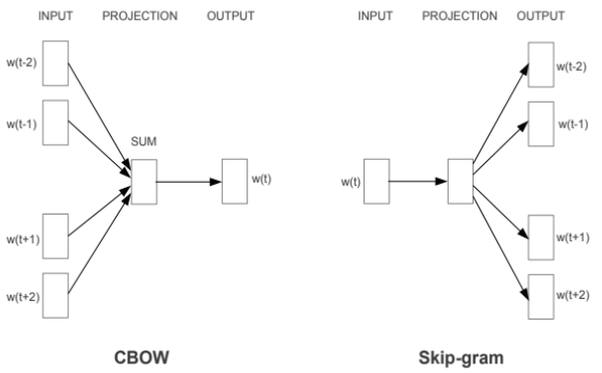

Fig. 1. CBOW, Skip-gram of Word2Vec

이런 특징을 가진 트랜스포머 모델은 병렬 처리 능력, 장거리 의존성에 대한 학습, 유연한 적용성으로 효과적으로 다양한 자연어처리 작업 수행에 적용되고 있다.

또한, 본 연구에서 학습에 사용할 임베딩도 일반 트랜스포머 모델이 사용하는 임베딩보다 낮은 차원으로 추출하여 사용하는데, Word2Vec 모델[12]을 통해 임베딩 추출을 진행한다. 이 모델은 데이터 내에서 단어의 의미를 벡터의 형태로 표현하는 기법으로, 단어를 고차원 벡터로 변환함으로써 단어 간의 의미적 유사성을 수치적으로 나타낼 수 있다. Word2Vec 모델은 크게 두 가지 접근 방식이 있는데, CBOW(Continuous Bag of Words)와 Skip-Gram 방식이 있다. Fig. 1과 같이 CBOW 방식은 주변에 있는 단어들을 사용하여 중심 단어를 예측하는 방식을 학습하고, Skip-Gram 방식은 CBOW 방식과는 반대로 중심 단어로부터 주변의 단어들을 예측하는 방식이다. 음절 단위의 형태소 분석을 진행하는 본 연구에서는 주변 음절을 통해 중심 음절의 태그를 예측하는 것과 비슷한 구조인 CBOW 방식을 활용하여 낮은 차원의 음절 단위 임베딩을 추출한다.

최근에도 딥러닝 기반 한국어 형태소 분석 관련 연구가 진행되고 있다. 한국어 형태소 분석과 품사 태깅을 위한 2단계 딥러닝 기반 파이프라인 모델을 소개한 연구[10]에서는 BERT 모델과 bi-LSTM 모델의 결합을 통해 시퀀스-투-시퀀스 방식으로 형태소 분리 및 음절 단위 품사 태그 부착을 진행하는 연구이다.

또한 BERT 모델과 트랜스포머 모델을 결합한 BERT-Fused 모델에 기반한 형태소 분석 기법[13]에 대한 연구도 소개되었다. 이 연구는 형태소 분석 작업을 형태소 분석을 어절 시퀀스를 형태소 시퀀스로 변환하는 기계 번역의 시점을 적용하여 관련 태스크에 높은 성능을 기록한 BERT-fused 모델을 사용하여 형태소 분석 작업을 수행한다.

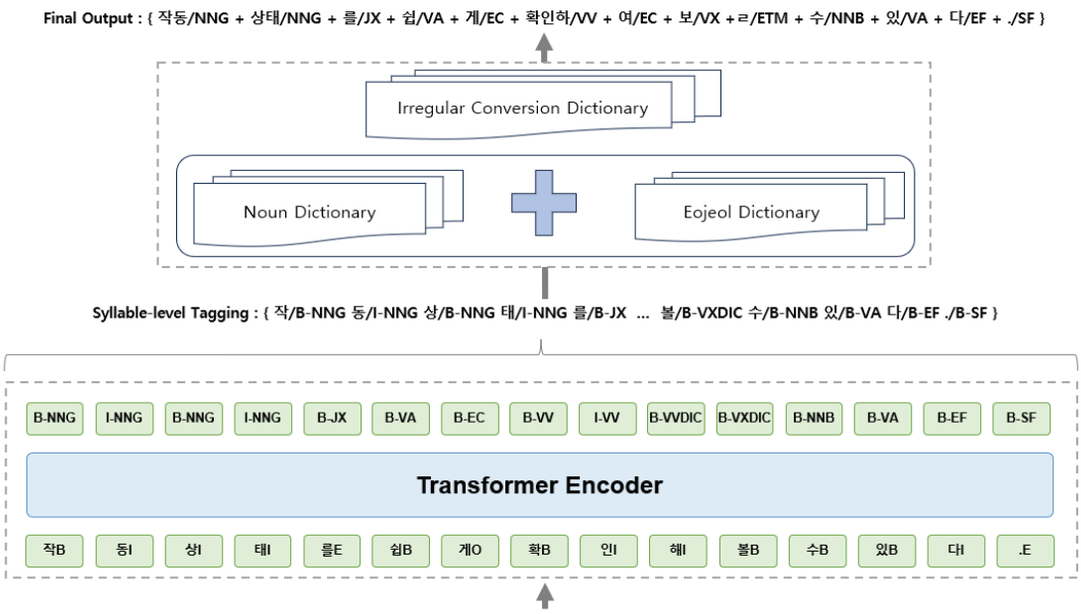

Fig. 2. Overall Diagram of the Morphological Analyzer

3. 연구 내용

3.1 음절 단위 임베딩 추출

일반 트랜스포머 관련 모델보다 높은 경량화를 달성하기 위해서 저차원의 임베딩 벡터가 필요하다. 이를 위해 본 연구에서는 다양한 주제, 어휘, 문체를 포함한 신문 기사 데이터를 사용했다. 신문 기사 데이터는 교정과 편집을 거쳐 문법적으로 정확하고 구조적으로 정돈된 텍스트를 제공하기 때문에 형태소 분석에 적합하다. 위와 같은 이유로 신문 기사 데이터를 토대로 음절 임베딩 추출을 진행하였다.

그러나 한국어의 특성상 음절 자체로는 충분한 의미 정보를 담지 못하기 때문에, 음절의 위치 정보를 나타내는 BIE 태그와 형태소 단위의 BI 태그를 추가하여 의미적 정보를 보완하였다. 또한, 음절 단위로 후보 품사 태그를 부착하는 과정을 통해 더 정확한 분석을 가능하도록 하였다.

임베딩 추출 단계에 사용된 데이터는 국립국어원에서 제공하는 모두의 말뭉치 서비스에서 2009년부터 2022년까지의 종합지, 전문지, 인터넷 기반 신문 매체의 기사를 포함한 신문 말뭉치 데이터셋이다. 이 데이터셋은 총 58,933,989개의 문장으로 구성되어 있으며, 기사 내용만 추출한 데이터의 크기는 15.54GB에 달한다.

이 신문 기사 데이터를 바탕으로 Word2Vec 모델의 CBOW 방식을 통해 27,562개의 음절에 대한 임베딩을 추출했다. 임베딩 벡터는 KorBERT의 임베딩 차원인 768차원보다 작은 128차원, 256차원, 512차원으로 설정하여 사용하였다.

3.2 음절 단위 품사 태깅

어절 단위 형태소 품사 태깅을 진행하기 위해서는 먼저 각 음절에 후보 품사 태그를 부착하는 순차적 레이블링 작업이 필요하다. 본 연구에서는 좀 더 정교한 음절 단위 품사 태깅을 위해, 기본 품사 태그 41개에서 몇 가지 요소를 추가했다. 구체적으로, 형태소 내의 위치를 나타내는 BI 태그와, 불규칙 변환이 필요한 형태소를 나타내는 DIC 태그를 추가하여 총 120개 후보 품사 태그로 확장해서 사용했다.

위 내용을 바탕으로 음절 단위 트랜스포머 인코더 모델의 학습을 진행하게 된다. Fig. 2의 하단 부분에서 보여주듯, 각 음절에 BIE 태그가 부착된 원본 문장이 모델의 입력으로 들어간다. 다음으로 Table 1에서 볼 수 있듯이, 입력된 문장에서는 각 음절마다 어절 단위의 BIE 태그, 형태소 단위의 BI 태그, 그리고 형태소 후보 품사 태그가 할당된다.

모델의 출력은 각 토큰 위치마다 120개의 후보 품사 태그에 대한 확률값을 계산한다. 이 확률값과 실제 정답 태그를교차 엔트로피(Cross Entropy) 손실 함수를 사용해 비교하여 모델을 학습한다. 학습이 진행되면서, 각 음절에 대해 가장 높은 확률을 가진 후보 품사 태그와 음절 쌍이 결정되며, 이는 다음 단계인 어절 단위 형태소 원형 복원 및 품사 태그 부착 단계의 입력으로 사용된다.

Table 1. Syllable-level Tag Types, Examples

	eojeol BIE	pos BI	syllable POS
작	B	B	B-NNG
동	I	I	I-NNG
상	I	I	B-NNG
태	I	I	I-NNG
를	E	I	B-JX

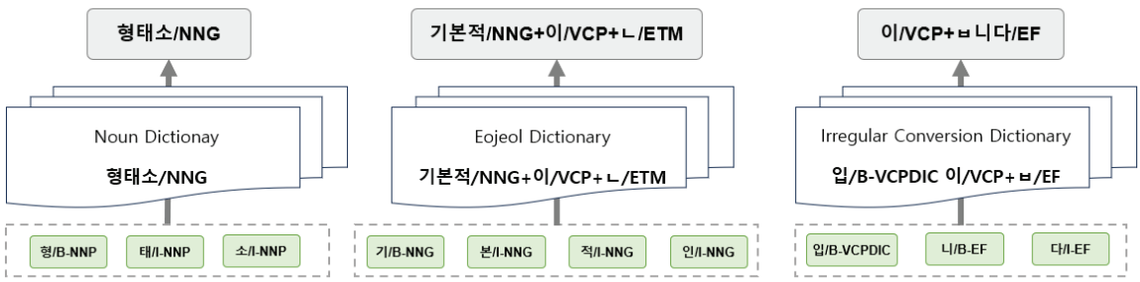

Fig. 3. Examples of Dictionaries's Usage

3.3 어절 단위 형태소 복원 및 품사 태깅

음절 단위로 후보 품사 태그가 부착된 결과를 바탕으로, 어절 단위의 형태소 원형 복원과 품사 태깅을 진행할 수 있다. Fig. 2의 윗부분에서 표현된 것처럼, 이 과정은 명사 사전, 어절 사전, 불규칙 변환 사전이라는 총 3가지 사전을 통해 진행된다. 먼저, 명사 사절과 어절 사전이 포함된 기분석 사전을 사용해 형태소 원형 복원 단계가 시작된다. 이 단계에서는 입력된 어절을 분석하여 해당 어절의 원형을 복원하고, 그 원형에 맞는 품사 태그를 부착한다. 불규칙 변환이 필요한 경우, 불규칙 변환 사전을 활용하여 형태소를 적절히 변환 후, 품사를 태깅한다.

명사 사전은 두 음절 이상의 명사들로 구성되어 있으며, 명사에 대한 품사 태깅 과정에서 중요한 역할을 한다. 명사 태깅 시, 해당 명사가 명사 사전에 등록되어 있는지 검색을 진행한 후, 검색된 명사의 품사와 입력된 태깅이 다를 경우에 사전에 저장된 품사로 수정하여 복원 작업을 진행한다. 예를 들어 Fig. 3처럼, 형태소의 각 음절이 NNP 태그로 입력되었더라도, 명사 사전에서 해당 명사를 검색했을 때 다른 품사 정보가 있다면, 그 정보에 따라 태깅이 수정된다. 이 과정은 태깅의 정확성을 높이고, 잘못된 태깅을 사전에 등록된 정보로 교정하는 데 중요한 역할을 한다.

어절 사전에서는 중의성을 해소하기 위해 해당 어절뿐만 아니라 이전 어절의 마지막 형태소와 다음 어절의 첫 형태소를 참고하여 어절이 사전에 등록되어 있는지 확인한다. 그 후, 사전에 구축된 품사 정보를 기반으로 해당 어절을 적절한 품사로 태깅한다. 이를 통해 어절 간 문맥을 고려하여 보다 정확한 품사 태깅을 가능하게 한다.

마지막으로 불규칙 변환 사전을 통해 불규칙 형태소를 처리한다. 음절 단위 품사 태깅 과정에서 불규칙 변환이 필요한 음절 단위 형태소 태그에는 DIC 태그가 부여된다. 이 DIC 태그가 붙은 음절들을 불규칙 변환 사전에 검색되어 적절한 형태소로 복원되며, 동시에 해당 품사가 태깅된다. 이러한 과정을 통해 어절 단위 형태소 복원 및 품사 태깅이 최종적으로 완료된다.

4. 실험 및 평가

4.1 데이터셋

실험 및 평가 단계에서 형태소 품사 태깅 학습 및 평가를 위해 세종 코퍼스를 사용했다. 실제 사용한 학습 데이터는 총 27,141문장에 1,180,565음절로 각 문장당 평균 43음절로 이루어져 있고, 평가 데이터는 총 6,702문장에 296,207음절로 각 문장당 평균 44음절이다.

4.2 실험을 위한 모델 매개변수 관련

본 연구는 트랜스포머 인코더 모델을 활용하여 형태소 분석 작업을 수행하였다. 그러나 다른 사전학습 모델에서 미세조정 작업을 수행한 것이 아닌, 경량화를 위해 임베딩 추출 단계에서부터 낮은 차원으로 추출하여 모델 학습을 진행하였다. 또한, 계층 개수마다 따로 모델을 학습하였다.이 차이점을 파라미터 수로 알아보기 위해 KorBERT와 각 계층 수 및 여러 임베딩 차원의 트랜스포머 인코더에 대한 파라미터 수 비율을 Table 2와 같이 비교하였다.

4.3 음절 단위 품사 태깅 평가

전체 단계에 대한 평가는 Table 3에 나와 있으며, 각 단계에서 정확도를 평가지표로 사용하였다. 음절 단위 품사 태깅 단계에서는 트랜스포머 인코더를 사용하여 각 음절에 후보 형태소 태그를 부착하는 작업이 진행된다. 여러 모델 중에서 가장 높은 성능을 기록한 모델은 계층 6개와 256의 임베딩 차원의 모델로, 음절 단위 정확도에서 97.06%의 정확도를 기록하였다.

4.4 어절 단위 형태소 복원 평가

이 단계는 음절 단위 품사 태깅 단계에서 실행된 결과를 바탕으로 진행되며, 어절 단위의 형태소 복원과 품사 태깅을 수

Table 2. Parameter Ratio of Each Parameter Model to KorBERT

n_layers \ emb_dim	128	256	512
6	5.38%	10.82%	21.60%
5	5.02%	10.09%	20.15%
4	4.66%	9.37%	18.71%
3	4.30%	8.65%	17.27%
2	3.94%	7.93%	15.83%
1	3.58%	7.21%	14.39%

Table 3. Syllable-level and Eojeol-level Accuracy by Number of Layers and Embedding's Dimension

n_layers \ emb_dim	Syllable-level Acc.			Eojeol-level Acc.		
	128	256	512	128	256	512
6	96.65%	**97.06%**	96.78%	95.85%	**96.37%**	96.02%
5	96.22%	96.30%	96.57%	95.34%	95.87%	95.84%
4	96.10%	96.34%	96.60%	95.24%	95.48%	95.76%
3	95.34%	96.25%	96.53%	94.31%	95.42%	95.61%
2	95.09%	94.82%	96.37%	94.11%	93.42%	95.41%
1	93.24%	94.46%	95.78%	92.31%	93.08%	94.72%

행한다. 어절 단위 정확도는 음절 단위보다 다소 낮은 경향이 있다. 이는 어절 단위 단계에서의 정확도는 어절 단위에서 모든 형태소가 원형으로 복원되어야 해당 어절이 제대로 복원되었다고 평가하기 때문에, 복잡성이 증가하여 이전 단계보다 정확도가 하락하는 경향이 있기 때문이다.

Table 3에서 기록된 바와 같이 음절 단위에서 정확도 97.06%를 기록한 모델을 토대로 진행했을 때, 96.37%의 어절 단위 정확도를 기록하였다. 이는 어절 단위 복원의 복잡성으로 인해 약간의 정확도 하락이 발생했음을 보인다.

4.5 모델 매개변수 간 성능 비교

Table 2는 트랜스포머 인코더의 계층 수와 임베딩 차원에 따른 KorBERT와의 파라미터 수를 비교한 결과이다. 기본적으로 임베딩 차원의 크기에 비례하여 파라미터 수도 동일한 배율로 증가한다. 계층 수와 임베딩 차원이 커질수록 모델의 복잡성과 파라미터 수가 증가하게 된다.

가장 높은 음절 및 어절 단위 정확도를 기록한 모델은 계층 6개와 256차원 임베딩을 사용하는 모델로, 음절 단위에서는 97.06%, 어절 단위에서는 96.37%의 정확도를 보인다.

본 연구에서 제안하는 모델은 음절 기반 단위 임베딩을 사용한다. 음절은 한국어에서 의미를 가지는 최소 단위인 형태소보다 낮은 수준의 단위이므로, 음절 자체가 가질 수 있는 정보량은 제한적이다. 이러한 이유로 512차원 임베딩보다 256차원 임베딩이 음절 단위의 정보 처리를 위한 더 적절한 선택으로 분석된다.

이 모델은 KorBERT의 파라미터 수의 10.82%만을 사용하여 경량화에 성공하였으며, 이는 다른 언어 모델과 병행하여 사용하기에도 충분한 효율성을 보여준다. 또한, 이 모델뿐만 아니라 다양한 매개변수를 가진 모델들 역시 어절 단위 정확도에서 최소 94%의 성능을 기록하였다. 이를 통해 상황에 따라 정확도와 모델 경량화 수준을 유연하게 선택할 수 있다.

가장 높은 성능을 보인 256차원 음절 임베딩을 기준으로 평가 데이터를 사용하여 추론 시간을 측정한 결과, 평가 데이터로 사용된 총 6,702개의 문장에 대해 음절 단위 품사 태깅을 진행하였으며, 각 실험을 10번씩 반복한 평균 시간은 표 4에 기록되어 있다.

Table 4에 기록된 추론 시간에 따르면, 6,702개의 문장에 대해 음절 단위 태깅을 완료하는 데 모든 계층에서 1.5초 이하의 시간이 소요되었다. 이는 bi-LSTM-CRFs[2] 모델이 같은 작업을 수행하는 데 걸리는 66.43초에 비해 약 44.5배 빠른 속도를 보여준다. 따라서 이 모델은 경량화뿐만 아니라 속도 면에서도 뛰어난 성능을 보임을 확인할 수 있다. 비록 KorBERT를 사용한 모델들[8, 10]에 대해서는 직접적인 시간 비교를 진행하지 못했지만, 본 연구에서 제안한 모델이 KorBERT에 비해 파라미터가 약 10배 적기 때문에, 더 빠른 추론 속도를 가질 것으로 예상한다.

Table 4. Average Inference Time at the Syllable-level by Number of Layers

n_layers	inference time
6	1.4901 sec
5	1.3903 sec
4	1.2652 sec
3	1.1425 sec
2	1.0322 sec
1	0.9234 sec

위와 같은 기록들을 토대로 경량화된 트랜스포머 인코더를 활용한 형태소 분석은 빠른 추론 속도를 보여주고 있고, 다른 언어 모델과 병행하여 실행이 가능한 경량화를 달성했음을 확인할 수 있다.

5. 결 론

본 연구는 트랜스포머 인코더를 활용한 음절 기반 형태소 분석기를 제안하였다. 신문 기사 데이터셋을 사용해 다양한 차원의 음절 임베딩을 추출하고, 이를 바탕으로 트랜스포머 인코더를 다양한 매개변수로 학습시켰다. 그 결과, 파라미터 수에 따른 경량화와 추론 속도를 평가할 수 있었다.

가장 높은 정확도를 기록한 모델은 ETRI의 KorBERT 대비 10.82%의 파라미터만을 사용하면서도, 음절 단위 정확도 97.06%와 어절 단위 정확도 96.37%를 달성했다. 이는 KorBERT를 사용한 연구[10]의 어절 단위 정확도 95.99%와 비교했을 때, 본 연구에서 제안한 형태소 분석기가 경량화와 함께 높은 정확도를 유지하고 있음을 보여준다.

본 연구는 음절 단위의 형태소 분석에서 트랜스포머 인코더의 적용 가능성을 입증하며, 적은 파라미터로도 높은 정확도를 유지할 수 있음을 제시한다. 이를 통해, 한국어 자연어 처리에서 경량화된 모델의 효율성과 실용성을 강조함으로써, 효율적인 형태소 분석 딥러닝 모델 개발에 대한 기여를 기대한다.

또한, 제안된 경량화 모델이 KorBERT 대비 적은 자원으로도 높은 성능을 낼 수 있다는 점에서, 실제 현장에서 대규모 데이터 처리가 필요한 응용 프로그램에 적합한 솔루션을 제공할 수 있다. 특히 자원 제한이 있는 환경에서 빠르고 효율적인 형태소 분석기를 적용함으로써, 자연어 시스템의 성능 최적화와 비용 절감에 기여할 수 있을 것이다.

하지만, 어절 단위 형태소 품사 태깅 단계에서는 트랜스포머 모델을 활용하지 않고 사전을 활용한 태깅으로 모델의 학습보다 사전의 구축 정도에 따라 정확도에 변동이 있을 수 있다. 이를 해결하기 위해 후속 연구에서는 후처리 단계까지 트랜스포머 모델을 활용하는 경량화된 시퀀스-투-시퀀스 모델을 탐색하고자 한다.

References

[1] Z. Huang, W. Xu, and K. Yu. "Bidirectional LSTM-CRF models for sequence tagging." *arXiv preprint arXiv:1508.01991*, 2015.

[2] H. M. Kim, J. M. Yoon, J. H. An, K. M. Bae, and Y. J. Ko, "Syllable-based Korean POS Tagging using POS Distribution and Bidirectional LSTM CRFs," in *Proceedings of the 28th Human & Cognitive Language Technology*, pp.3-8, 2016. (in Korean)

[3] X. Ma and E. Hovy, "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Vol.1, pp.1064-1074, 2016.

[4] 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 Korean Institute of Information Scientists and Engineers*, Vol.45, No.8, pp.792-800, 2018. (in Korean)

[5] E. I. Chung, and J. G. Park, "Word Segmentation and POS tagging using Seq2seq Attention Model," in *Proceddings of the 28th Human & Cognitive Language Technology*, pp.217-219, 2016. (in Korean)

[6] 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," *Journal of Korean Institute of Information Scientists and Engineers*, Vol.44, No.1, pp.57-62, 2017. (in Korean)

[7] J. W. Min, S. H. Na, J. H. Shin, and Y. K. Kim, "End-to-End Neural Transition-based Morpheme Segmentation and POS Tagging of Korean," in *Proceedings of the Korean Information Science Society Conference*, pp.566-568, 2019. (in Korean)

[8] Y. S. Choi and K. J. Lee, "Performance Analysis of Korean Morphological Analyzer based on Transformer and BERT," *Journal of Korean Institute of Information Scientists and Engineers*, Vol.47, No.8, pp.730-741, 2020. (in Korean)

[9] B. S. Choe, I. H. Lee, and S. G. Lee, "Korean Morphological Analyzer for Neologism and Spacing Error based on Sequence-to-Sequence," *Journal of Korean Institute of Information Scientists and Engineers*, Vol.47, No.1, pp.70-77, 2020. (in Korean)

[10] J. Y. Youn and J. S. Lee, "A Deep Learning-based Two-Steps Pipeline Model for Korean Morphological Analysis and Part-of-Speech Tagging," *Journal of Korean Institute of Information Scientists and Engineers*, Vol.48, No.4, pp.444-452, 2021. (in Korean)

[11] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, Vol.30, 2017.

[12] T. Mikolov et al., "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[13] C. J. Lee and D. Y. Ra, "Korean Morphological Analysis Method Based on BERT-Fused Transformer Model," *THE KIPS Transactions on Software and Data Engineering*, Vol.11, No.4, pp.169-178, 2022. (in Korean)

[14] KorBERT [Internet], https://aiopen.etri.re.kr/bertModel

민 수 영

https://orcid.org/0009-0002-2591-2578
e-mail : sujae9704@gmail.com
2023년 성균관대학교 소프트웨어학과(학사)
2023년 ~ 현 재 성균관대학교
AI시스템공학과 석사과정
관심분야 : 자연어처리, 텍스트 요약

고 영 중

https://orcid.org/0000-0002-0241-9193
e-mail : youngjoong.ko@gmail.com
2004년 ~ 2019년 동아대학교
컴퓨터공학과 교수
2019년 ~ 현 재 성균관대학교
소프트웨어학과 교수
관심분야 : 자연어처리, 정보검색, 언어모델, 생성형 AI, 대화 시스템, 질의응답 시스템, 문서 분류 및 요약 시스템 등