Journal of The Korea Convergence Society
Vol. 12. No. 12, pp. 39-47, 2021

ISSN 2233-4890 / e-ISSN 2713-6353
DOI : https://doi.org/10.15207/JKCS.2021.12.12.039

한국어 단어 및 문장 분류 태스크를 위한 분절 전략의 효과성 연구

김진성[1], 김경민[2], 손준영[2], 박정배[3], 임희석[4*]
[1]고려대학교 컴퓨터학과 석사과정, [2]고려대학교 컴퓨터학과 석박사통합과정,
[3]고려대학교 Human-inspired AI연구소 연구교수, [4*]고려대학교 컴퓨터학과 교수

A Comparative study on the Effectiveness of Segmentation Strategies for Korean Word and Sentence Classification tasks

Jin-Sung Kim[1], Gyeong-min Kim[2], Jun-young Son[2], Jeongbae Park[3], Heui-seok[4*]
[1]Master Student, Department of Computer Science and Engineering, Korea University
[2]Master & Ph. D. Combined Student, Department of Computer Science and Engineering, Korea University
[3]Research Professor, Human-inspired Computing Research Center, Korea University
[4]Professor, Department of Computer Science and Engineering, Korea University

요 약 효과적인 분절을 통한 양질의 입력 자질 구성은 언어모델의 문장 이해력을 향상하기 위한 필수적인 단계이다. 입력 자질의 품질 제고는 세부 태스크의 성능과 직결된다. 본 논문은 단어와 문장 분류 관점에서 한국어의 언어적 특징을 효과적으로 반영하는 분절 전략을 비교 연구한다. 분절 유형은 언어학적 단위에 따라 어절, 형태소, 음절, 자모 네 가지로 분류하며, RoBERTa 모델 구조를 활용하여 사전학습을 진행한다. 각 세부 태스크를 분류 단위에 따라 문장 분류 그룹과 단어 분류 그룹으로 구분 지어 실험함으로써, 그룹 내 경향성 및 그룹 간 차이에 대한 분석을 진행한다. 실험 결과에 따르면, 문장 분류에서는 자모 단위의 언어학적 분절 전략을 적용한 모델이 타 분절 전략 대비 최대 NSMC: +0.62%, KorNLI: +2.38%, KorSTS: +2.41% 높은 성능을, 단어 분류에서는 음절 단위의 분절 전략이 최대 NER: +0.7%, SRL: +0.61% 높은 성능을 보임으로써, 각 분류 그룹에서의 효과성을 보여준다.

주제어 : 언어학적 분절, 문장 분류, 단어 분류, 토큰화, 자연어 처리, 사전학습 언어모델

Abstract The construction of high-quality input features through effective segmentation is essential for increasing the sentence comprehension of a language model. Improving the quality of them directly affects the performance of the downstream task. This paper comparatively studies the segmentation that effectively reflects the linguistic characteristics of Korean regarding word and sentence classification. The segmentation types are defined in four categories: eojeol, morpheme, syllable and subchar, and pre-training is carried out using the RoBERTa model structure. By dividing tasks into a sentence group and a word group, we analyze the tendency within a group and the difference between the groups. By the model with subchar-level segmentation showing higher performance than other strategies by maximal NSMC: +0.62%, KorNLI: +2.38%, KorSTS: +2.41% in sentence classification, and the model with syllable-level showing higher performance at maximum NER: +0.7%, SRL: +0.61% in word classification, the experimental results confirm the effectiveness of those schemes.

Key Words : Linguistic segmentation, Sentence classification, Word classification, Tokenization, Natural language processing, Pre-trained language model

*This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).
*Corresponding Author : Heuiseok Lim(limhseok@korea.ac.kr)
Received October 25, 2021 Revised November 9, 2021
Accepted December 20, 2021 Published December 28, 2021

1. 서론

문장의 분절, 즉 우리가 문장을 어떻게 끊어 읽는지는 주어진 문장에 대한 이해와 직결된다. '아버지가 방에 들어가신다'와 '아버지 가방에 들어가신다', 문장을 분절하는 방식에 따라 그 의미가 전혀 달라진다. 단어 혹은 문장을 적절히 분류하는 자연어 처리 세부 태스크의 수행에서도, 주어진 문장의 적절한 분절을 통해 양질의 입력 자질을 구성하는 것은 언어모델이 문장의 의미를 효과적으로 파악하는 데에 직접적인 영향을 미친다. 즉, 적절한 분절 기법의 적용으로 입력 자질의 품질을 제고하는 것은 세부 태스크의 최종 성능도 향상 시킬 수 있음을 의미한다. 따라서, 양질의 입력 자질 구성을 위해, 문장을 적절히 분절하는 전략을 선택하는 작업은 필수적이다. 본 연구에서의 적절한 분절 전략은, 곧 분절 시에 해당 언어의 특징을 효과적으로 반영함을 일컫는다. 세부 태스크에서 대상 언어의 특징을 잘 고려하기 위해 다양한 기법을 사용할 수 있으나, 본 연구에서는 문장의 분절 관점에서 한국어의 특징을 효과적으로 반영 가능한 분절 전략을 찾아내기 위한 비교 검증에 집중한다. 한국어의 특징을 효과적으로 반영하는 분절 전략을 찾기 위해서는, 먼저 한국어가 가진 특징 중 입력 자질의 품질에 영향을 미칠 수 있는 특징에는 무엇이 있는지에 대한 언급이 필요하다. 특히, 한국어가 가지는 교착어적 특성 및 음소문자로서의 특성은 주어진 문장의 의미를 파악하여 문장 혹은 단어를 분류하는 세부 태스크에서 어려움으로 작용한다.

문장의 의미 파악을 어렵게 만드는 한국어의 특징은 크게 두 가지를 고려할 수 있다. 첫 번째로, 한국어는 알파벳을 사용하는 영어나 라틴계 유럽어와 같은 굴절어와는 달리 교착어적 특성을 가진다. 접사의 역할을 하는 조사 혹은 어미가 어근과 결합하여 어절의 의미를 형성한다는 것이다. 이로 인해 조사나 어미가 변함에 따라 어절의 의미 역시 달라지는데, 접사에 의해 서술어의 역할이 능동, 피동 혹은 사동으로 변화하기도, 문장 내의 주어와 목적어 간의 관계가 바뀌기도 한다. 예를 들어, '덜미를 잡았다.'와 '덜미를 잡혔다.'라는 두 문장은 선어말 어미 '-히-'의 존재로 인해 문장의 의미 구조 즉, 능동과 피동의 여부가 달라진다. 또한, '나'라는 체언 뒤에 '-는'이 오는지 혹은 '-를'이 오는지에 따라, 어절의 역할이 주어 '나는'인지, 목적어 '나를'인지가 좌우된다.

두 번째로, 한국어는 음소문자로서 형태소 혹은 음절 하나가 초성, 중성 그리고 종성의 결합으로 이뤄진다는 점이다. 이는 이미 문자 하나가 최소의 단위이자 완성된 형태로 존재하는 알파벳 언어 혹은 같은 교착어인 일본어나 중국어와 달리, 한국어에서는 자모 체계에 따른 음절의 의미 변화까지 고려되어야 함을 의미한다. 예를 들어, '강'의 초성이 변화하여 '상'이 되거나, 중성이 변하여 '공' 혹은 종성이 변하여 '감'이 되면 모두 원래의 음절인 '강'의 의미는 사라지고, 다른 의미로 쓰이게 된다. 이렇게 자모의 변화에 따라 의미가 달라진 음절은 또 다른 음절과 결합하여 각각 '공상하다', '공공하다' 혹은 '공감하다'로 구성될 때, 동사의 의미가 서로 달라진다.

전통적으로 자연어 처리에서의 분류 태스크는, 분류 단위에 따라 단어 분류 태스크와 문장 분류 태스크로 나뉜다. 전자는 주변 문맥을 고려하여 단어가 속하는 클래스를 예측하는 태스크이며, 후자는 문장의 종합적 의미를 이해하고 추론하는 태스크이다. 한 가지 주목할 점은, 이러한 태스크의 구분에 따라 분절 전략이 다른 효과를 가질 수 있다는 점이다. 예를 들어, 문장 '우리의 밤은 아름답다'에서 '밤'이라는 단어의 의미를 시간인지 열매인지 분류하는 작업에서는 해당 단어를 문법 형태소 '-은'으로부터 반드시 분리가 필요한 반면, 문장의 감정이 긍정적인지 부정적인지 분류하는 문장 분류 작업이라면 '밤'이라는 단어를 정확히 분리할 필요가 없기 때문이다. 따라서, 우리는 검증을 위한 세부 태스크를 문장 그룹과 단어 그룹, 두 그룹으로 나누어, 단어 그룹에는 개체명 인식, 의미역 결정, 의존 구문 분석 태스크를, 문장 그룹에서는 감성 분석, 자연어 추론, 문장 간 의미적 유사도 판단 태스크를 검증한다. 이를 통해 분절 전략에 따른 그룹 내에서의 경향성 분석 및 그룹 간의 차이에 대한 비교를 진행한다. 자연어 처리 세부 태스크의 성능과 직결되는 입력 자질의 품질 향상을 위한 분절 전략 연구의 필요성에도 불구하고, 효과적인 한국어 분절에 대한 객관적 결과 비교와 함께 언어학적 분석을 제시하는 종합적 연구는 아직까지 미흡하다. 또한, 효과적인 분절 전략이 각기 다를 수 있는 단어 및 문장 태스크 그룹 간의 비교 연구는 존재하지 않는다.

본 논문에서는 첫 번째로, 각 분절 전략에 따른 성능 결과의 제시를 통해, 한국어 분절에 대한 객관적 지표를 제공한다. 두 번째로, 각 분절 전략에 대한 언어학적 관점의 분석을 통해, 한국어에서 언어학적 분절의 중요성에 대해 강조한다. 마지막으로, 단어 분류 그룹과 문장 분류 그룹 간 분절 전략에 따른 효과성의 차이를 비교 검증한다.

본 논문의 구성은 다음과 같다. 2장에서는 분류 단위에 따른 언어학적 분절의 필요성과 관련 연구를 소개하고, 3장에서는 언어학적 단위의 분절 전략을 네 가지로 나누어 제시한다. 4장에서는 세부 태스크 및 실험에 관해 설명하고, 5장에서 결론을 짓는다.

2. 관련 연구

사전학습 언어모델 최신의 자연어처리 분야에서는 다양한 세부 태스크에서 뛰어난 성능을 보여주는 사전 학습 언어모델의 활용이 선호된다. 본 연구의 수행을 위한 모델의 사전 학습에 사용된 기반 모델 구조는 페이스북(Facebook)에 의해 제안된 RoBERTa(Robustly Optimized BERT pre-training approach)[1]이다. 이 모델은 BERT[2]와 동일한 구조를 사용하지만, BERT의 학습 기법 중 문장 간 순서를 예측하는 NSP(Next Sentence Prediction) 기법을 배제하고, 무작위로 15%의 토큰을 정적으로 마스킹하는 정적 마스킹 기법을 동적 마스킹(dynamic masking)으로 대체하여 학습을 진행하였다.

분절 관련 기존 연구 자연어 처리 분야의 고질적인 어려움 중 하나인 OOV(Out-Of-Vocabulary) 문제를 해결하기 위하여, 단어를 보다 더 작은 단위인 서브단어 단위로 분절하여 토큰화 하는 기법인 서브워드 알고리즘이 출현하였으며, 대표적인 서브워드 알고리즘으로는 BPE[3], Wordpiece[4], Sentencepiece[5]가 있다. 이처럼 한국어 분야에서도 OOV 문제를 돌파하기 위한 연구에서 시작하여 제안된 분절 기법에 관한 연구들이 존재한다[6-8]. 토큰화 단계의 효과적인 구성을 위하여 개별 분절 기법의 적용을 시도한 연구들도 있는데, 자모 단위의 토큰화가 OOV 문제의 발생 확률을 낮추는 데 효과적이라고 검증한 연구[9], 한국어 어절의 분해 혹은 음절의 분해 및 재조립에 주목한 연구[10,11] 등이 이에 해당한다. 또한, 여러 개별 단위의 토큰화 기법을 구성한 연구가 존재한다[12,13].

하지만, 이러한 개별 연구들의 결과들로는 한국어 분류 태스크에서 분절 전략 간의 성능 차이만을 객관적으로 비교하기 어렵다. 즉, 세부 태스크들에서 분절 과정을 제외한 타 조건을 통일하여 분절 전략 간 비교 검증만을 진행함과 동시에, 분절 전략에 따른 태스크의 성능 결과를 언어학적 관점의 질적 분석을 통해 현상 규명 및 연구자의 이해를 돕는 종합적인 연구는 존재하지 않는다. 이로 인해 한국어 분류 태스크 연구자들이 입력 자질을 구성할 때, 효과적인 분절 전략에 관한 객관적 비교 지표 및 언어학적 분석의 부재로 인해 어려움을 겪을 수 있다. 이러한 측면에서 한국어 분류 태스크에서 분절 전략 간 명확한 효과성 비교에 대해 재고할 필요가 있다.

Table 1. The segmentation results of the example sentence '세종은 조선의 국왕이다.'

Input sentence	
세종은 조선의 국왕이다	
Segmentation	Segmentation results
Eojeol	세종은/조선의/국왕이다/.
Morpheme	세종/은/조선/의/국왕/이/다/.
Syllable	세/종/은/ /조/선/의/ /국/왕/이/다/.
Subchar	ㅅㅔㅈㅗㅇㅗㄴ/.../ㄱㅜㄱㅇㅏㅇㅇㅣㄷㅏ/.

3. 한국어 분류를 위한 언어학적 분절 전략

3.1 언어학적 분절 기법

한국어 문장의 의미를 잘 이해하고 적절한 단위로 분류하기 위해서는, 앞서 언급했듯 한국어의 두 가지 특징을 효과적으로 반영해야 한다. 한국어가 가지는 교착어적 특성, 즉 조사와 어미라는 문법 요소에 따른 어절의 의미 변화를 잘 포착하고, 또한 음소문자의 특성, 즉 자모의 결합에 따른 음절 혹은 형태소의 의미 변화를 적절히 고려해야 한다. 이렇듯 한국어에서는 작은 언어학적 단위의 차이가 어절 및 전체 문장의 의미 구조까지 변화시킬 수 있으므로, 이러한 특징의 효과적인 반영을 위해서는 영어와 같이 단어 단위의 분절이 아닌, 보다 언어학적으로 세분화된 언어학적 분절 전략의 적용이 필수적이다.

본 연구에서 '언어학적 분절' 유형의 범주는 언어학적 관점에서 네 가지 단위로 나누어, Table 1의 예시와 같이 각각 어절 단위, 형태소 단위, 음절 단위, 자모 단위로 구분한다. 각 분절 유형 중, 한국어 특성을 효과적으로 고려하여 가장 뛰어난 성능을 보이는 분절 전략을 검증한다. 또한, 네 가지 언어학적 분절과 함께 토큰화 단계에서 기존 연구를 통해 효과성이 입증된 서브워드 알고리즘[14]을 적용한다. Sentencepiece 알고리즘은 언어에 강건한 토큰화 기법으로, 교착어적 특성을 갖는 중국어, 일본어 그리고 한국어 즉, CJK(Chinese, Japanese and Korean)에 적합하다는 사전 연구 결과가 있다[5].

따라서 본 연구에서는 Sentencepiece 서브워드 알고리즘을 채택한다. Table 1은 각 언어학적 분절 전략에 따른 예시 문장 '세종은 조선의 국왕이다.'의 분절 결과를 보여주며, 서브워드 알고리즘을 통해 토큰화 하기 직전의 상태이다.

어절 단위 어절 단위로 분절한 뒤 서브워드 알고리즘을 적용한다. 기존의 서브워드 알고리즘을 한국어에 적용하는 방법과 동일하다. Table 1의 Eojeol 행에 해당한다.

형태소 단위 형태소 분석을 통해서 실질 형태소로부터 조사, 어미 등의 다양한 문법 요소를 분절한 뒤 서브워드 알고리즘을 적용한다. 즉, 형태소 분석기를 활용하여, 뜻을 가진 가장 작은 단위인 형태소로 분리한 후에 서브워드 알고리즘을 적용하는 방식이다. 형태소 단위의 문법적 정보를 사전에 제공한 뒤 모델을 학습하는 것의 효과성을 확인하고자 한다. Table 1의 Morpheme 행에 해당한다.

음절 단위 음절 단위로 분절한 뒤 서브워드 알고리즘을 적용한다. 단, 띄어쓰기도 하나의 토큰으로 간주한다. 예를 들어, '오늘 날씨가 참 좋네요.'라는 문장은 ['오', '늘', ' ', '날', '씨', '가', ' ', '참', ' ', '좋', '네', '요', '.']로 분절한다. 각각의 음절을 하나의 인코딩 단위로 보는 것이 어떤 영향을 주는지 분석하고자 한다. Table 1의 Syllable 행에 해당한다.

자모 단위 자모 단위 분절은 각 음절을 초성, 중성, 종성으로 분리한 뒤 서브워드 알고리즘을 적용한다. 이 기법을 통해 모델이 자모 체계를 학습하는 것이 한국어 분류 태스크에서 어떤 효과를 가지는지 분석한다. Table 1의 Subchar 행에 해당한다.

Fig. 1은 본 논문에서 실험한 전체 모델의 구조도를 보여준다. '세종은 조선의 국왕이다.'와 같은 입력 문장은 각각의 언어학적 분절 단위 즉, 어절, 형태소, 음절, 자모로 분절된 후 서브워드 알고리즘 Sentencepiece의 적용을 통해 토큰화되어 모델의 입력 자질로 구성된다.

이러한 각각의 언어학적 분절 기법이 적용된 네 가지 유형의 입력 자질은, RoBERTa 구조에 각 언어학적 분절 전략을 통해 사전학습된 모델에 학습 과정을 거쳐, 단어 분류 혹은 문장 분류 세부 태스크를 수행한다.

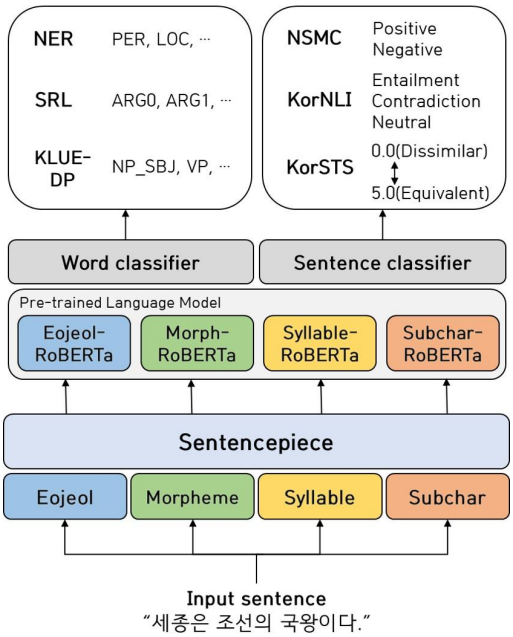

Fig. 1. The model architecture based on linguistic segmentation strategies

3.2 분절에 따른 단어 및 문장 분류에서의 효과성

네 가지의 언어학적 분절 중 어느 전략이 각각 단어 그룹과 문장 그룹에서 가장 효과적인지에 대답하고자, 분류 그룹별로 각 전략에 대한 언어학적 관점에서의 고찰을 진행한다.

단어 분류 태스크 단어 분류 태스크는 개체명 인식이나 의미역 결정과 같이 체언 중심의 분류 태스크와, 의존 구문 분석, 관계 추출 등의 서술어도 중요한 역할을 하는 태스크로 나누어진다. 먼저, 체언 위주의 태스크에서는 각 분절 전략이 주어진 어절 안에서 유의미한 형태소 즉, 체언을 다른 문법 요소로부터 정확히 분리해낼 수 있는지에 대한 여부가 태스크의 성능을 좌우한다고 할 수 있다. 특히, 유의미한 정보를 가진 개체(entity)를 적절한 태그로 분류하는 태스크의 경우가 이에 해당한다. 예를 들어, 표 1에서 입력 문장에 포함된 개체는 체언인 '세종'과 '조선'이다. 두 개체를 정확한 태그인 '인간(PER)'과 '위치(LOC)'로 분류하기 위해서는, 어절 '세종은'과 '조선의'로부터 조사 '-은'과 '-의'를 정확하게 분리하여 개체 정보인 '세종', '조선'만을 분류 작업에 사용할 수 있어야 한다.

또한, 체언뿐만 아니라 문장 내 전체 문장 성분 간

관계의 포착이 필요한 경우에는, 체언과 조사의 분리와 더불어 서술어의 의미 변화의 효과적인 파악도 요구된다. 예를 들어, 동사 '잡았다'와 '잡혔다'는 선어말어미 '-히-'의 존재로 인해 각각 취하는 문장 성분이 '범인을', '범인이'와 같이 목적어와 주어로 달라진다. 즉, 문장 분류 태스크와 같이 문장 성분 간의 의미 구조의 파악도 요구된다고 할 수 있다. 결국, 단어 그룹에서는 실질 형태소로부터 문법 형태소의 분리가 핵심인 교착어적 특성을 효과적으로 고려하는 분절 전략이나 서술어의 의미 변화를 반영 가능한 전략이 유리함을 알 수 있다.

문장 분류 태스크 문장을 의미에 따라 분류하기 위해서는 문장의 주요 성분인 주어, 서술어, 목적어를 파악하는 것이 중요하다. 그중 한국어는 핵말언어로서, 문장의 핵이 되는 서술어의 변화를 세밀하게 포착 가능한 분절 전략이 태스크에서 그 효과성을 보일 수 있다. 예를 들어, 문장이 능동문인지 피동문인지 여부, 문장의 시제가 현재인지 과거인지 등 문장 전체의 의미 구조를 파악하는 것이 중요하다는 것이다. 또한, 체언 혹은 용언으로부터의 문법 요소 분리는 전체적 의미 파악에서 중요하지 않은데, 이는 개체 혹은 문장 성분을 하나하나 정확히 태깅하는 성격의 작업이 아니기 때문이다. 결국, 서술어의 통사적 변화를 가능한 가장 작은 단위까지 고려하는 분절 전략이 태스크에서도 효과적임을 알 수 있다.

4. 실험

4.1 데이터셋

4.1.1 단어 분류 데이터셋

창원대x네이버 NER/SRL 데이터셋 본 데이터셋은 공개 챌린지의 목적으로 창원대에서 구축한 개체명 인식[1], 의미역 결정[2] 데이터셋으로, CoNLL[15] 포맷에 따라 주석이 달렸다. NER 데이터셋을 통해 주어진 문장에서 개체를 인식하여 알맞은 개체 유형으로 분류하는 개체명 인식 태스크를 수행하며, SRL 데이터셋을 통해 입력 문장에서 의미역을 찾아, 적절한 의미역 유형에 할당하는 의미역 결정 태스크를 수행한다. 각각의 말뭉치의 크기는 약 90K, 35K이다. 두 데이터셋 모두 평가 점수로 F1 점수(%)를 사용하였다.

KLUE-DP 데이터셋 KLUE 데이터셋[16]은 의존 구문 분석을 비롯한 8가지 한국어 자연어 이해 태스크를 위한 벤치마크 데이터셋이다. KLUE-DP는 9개의 구문 태그와 6개의 기능 태그로 구성된 TTA 의존 구문 태그[3] 기법에 따라 구성되었다. 해당 데이터셋을 통해 각 어절 간의 지배 관계를 분류하는 의존 구문 분석 태스크를 수행한다. 말뭉치의 크기는 총 14,000 문장으로, 10,000 문장의 학습 데이터 및 4,500 문장의 검증 및 평가 데이터로 이루어진다. 평가를 위해 주어진 어절의 지배소(head)를 예측하는 UAS(Unlabeled Attachment Score) 점수를 측정하였다.

4.1.2 문장 분류 데이터셋

네이버 NSMC 데이터셋 주어진 하나의 문장을 긍정적인 피드백인지, 부정적인 피드백인지 밝혀내는 태스크인 감성분석 태스크를 수행하기 위해 한국어로 공개된 네이버 영화 리뷰 데이터셋[4]이다. 입력된 문장을 0에 해당하는 부정적(Negative), 1에 해당하는 긍정적(Positive) 두 가지 표지로 분류한다. 150,000개의 훈련 데이터와 50,000개의 평가 데이터로 이루어져 있다. 성능 측정을 위해 정확도(Accuracy, %)를 활용하였다.

KorNLI 데이터셋 영어로 된 SNLI, MNLI, XNLI 데이터셋을 기계 번역하여 구성한 한국어 자연어 추론 데이터셋이다[17]. 주어진 두 문장 간의 관계를 함의(entailment), 모순(contradiction), 중립(neutral) 세 가지 유형으로 의미 관계를 분류하는 태스크인 자연어 추론 태스크에 활용한다. 말뭉치는 7.5K 개의 평가 예제를 포함한 95K 개의 예제로 이루어진다. NSMC 데이터셋과 마찬가지로 정확도(Accuracy, %)를 평가하였다.

KorSTS 데이터셋 KorNLI 데이터셋과 동일한 논문에서 공개된 데이터셋[17]으로, 해당 데이터셋은 텍스트 간의 의미적 유사성을 평가하는 태스크에 활용하며, 주어진 두 문장 간의 의미적 유사성의 정도를 0.0에서 5.0 사이의 실수값으로 산출한다. 0.0에 가까울수록 불일치, 5.0에 가까울수록 일치함을 나타낸다. 말뭉치의 크기는

1) https://github.com/naver/nlp-challenge/tree/master/missions/ner

2) https://github.com/naver/nlp-challenge/tree/master/missions/srl

3) https://aiopen.etri.re.kr/data/003.%EC%9D%98%EC%A1%B4%EA%B5%AC%EB%AC%B8%EB%B6%84%EC%84%9D_%EA%B0%80%EC%9D%B4%EB%93%9C%EB%9D%BC%EC%9D%B8.pdf

4) https://github.com/e9t/nsmc

5,749개의 훈련 예제와 2,879개의 평가 예제로 구성되었다. 평가를 위해 100 * 스피어맨 상관 계수(Spearman correlation)를 사용하였다.

Table 2. The hyperparameters in the fine-tuning

Hyper-parameters	NAVER-NER	NAVER-SRL	KLUE-DP
Max. epochs	10	5	10
Learning rate	{1e-5, 3e-5, 5e-5}		
Train batch size	128		
Test batch size	64		
Max seq. length	128(*256 for Syllable)		
Weight decay	0.1		
Hyper-parameters	**NSMC**	**KorNLI**	**KorSTS**
Max. epochs	5	10	15
Learning rate	{1e-5, 3e-5, 5e-5}		
Train batch size	128		
Test batch size	64		
Max seq. length	128(*256 for Syllable)		
Weight decay	0.1		

4.2 실험 환경

사전학습을 진행한 모델은 BERT 모델[2]의 학습 기법 중 문장 간 순서를 예측하는 NSP 기법을 제외하고, 정적으로 마스킹 된 토큰을 예측하는 MLM 기법을 동적 마스킹으로 대체한 사전학습 언어모델인 RoBERTa 모델[1]이다. RoBERTa 모델의 구조를 바탕으로, 한국어 위키피디아 데이터셋[5]으로 사전학습을 진행하였다. 사전학습 시의 하이퍼파라미터로는 Batch size는 8K, Learning rate은 1e-6, Max seq. length는 256, Total update steps)는 100K, Warmup steps는 6K로 설정하였다. 본 연구를 위한 RoBERTa 모델의 사전학습을 위해, FAIRSEQ[18] Pytorch 딥러닝 라이브러리[6]를 사용하여 RTX A6000 GPU * 4 으로 10일 동안 학습을 진행하고, 미세 조정을 위한 실험에는 단일 GPU로 학습을 진행하였다.

각 태스크를 위한 미세 조정 시에 설정한 하이퍼파라미터 값들은 Table 2를 통해 확인할 수 있다. 공통적으로는 사전학습을 통해 생성된 네 가지 모델을 각각 {1e-5, 3e-5, 5e-5}, 세 가지의 학습률과 네 가지의 랜덤 시드의 조합을 통해 실험을 진행하였다. 다시 말해서 하나의 분절 전략이 적용된 모델 하나당 총 12회의 반복 실험 성능을 측정하였다. 이때 Train batch size는 128, Test batch size는 64, Weight decay는 0.1로 설정하였다. Max seq. length는 128로 설정하되, 음절 단위 기법의 경우 한 음절씩 분리하므로 문장 내 정보 소실 방지를 위해 256으로 설정하였다. Max. epochs는 각각의 데이터셋에 따라 네이버 NER 데이터셋은 10회, 네이버 SRL 데이터셋은 5회, KLUE-DP 데이터셋에서는 10회 그리고 NSMC 데이터셋에서는 5회, KorNLI 데이터셋은 10회, KorSTS 데이터셋은 15회로 설정하였다.

Table 3. The performance results on each dataset based on segmentation scheme.

Segmentation strategy	Word classification		
	NAVER-NER	NAVER-SRL	KLUE-DP
Eojeol	85.99	78.08	90.99
Morpheme	85.34	77.59	89.81
Syllable	**86.04**	**78.20**	89.85
Subchar	86.03	78.17	**91.04**
Segmentation strategy	**Sentence classification**		
	NSMC	**KorNLI**	**KorSTS**
Eojeol	89.66	80.23	78.57
Morpheme	89.13	79.64	78.49
Syllable	89.36	77.89	77.11
Subchar	**89.75**	**80.27**	**79.52**

4.3 실험 결과 분석

Table 3은 언어학적 분절 전략에 따른 각 세부 태스크의 성능 결과를 보여준다. 표의 상단은 개체명 인식, 의미역 결정, 의존 구문 분석 태스크로 이루어진 단어 분류 그룹, 하단은 감성 분석, 자연어 추론, 문장 간 의미적 유사도 판단 태스크로 이루어진 문장 분류 그룹의 결과이다.

4.3.1 단어 분류 태스크

성능 분석 단어 분류 태스크들로 구성된 단어 그룹에서는 두 가지 경향성을 보임을 Table 3의 실험 결과를 통해 확인할 수 있다. 개체 혹은 의미역에 해당하는 체언의 분리가 필요한 개체명 인식, 의미역 결정 태스크에서는 음절 단위의 분절 전략을 적용한 모델이 우수한 성능을 보여준다. 음절 단위 전략은 형태소 단위의 분절과 비교하여 최대 +0.7%의 성능 차이를 보이며, 어절 단위 분절보다도 다소 높은 성능을 보인다. 한편, 단어 그룹과 문장 그룹의 중간적 성격을 띠는 의존 구문 분석 태스크

5) https://dumps.wikimedia.org/kowiki/latest/kowiki-latest-pages-articles.xml.bz2

6) https://github.com/pytorch/fairseq

의 경우 자모 기반의 분절 전략을 적용할 때 가장 높은 성능을 보인다. 형태소 단위의 분절 전략은 평균적으로 타 전략들에 비해 낮은 점수를 기록하였다.

언어학적 분석 형태소 단위의 분절은 실질 형태소로부터 문법 형태소를 분리하여 고려하지만, 형태소 분석기에 의존하므로 잘못된 분석 결과로 입력 자질을 구성할 경우 오히려 성능에 해가 될 우려가 있다. 자모 단위의 분절은 '-ㅂ고'와 같이, 어절 내 문법 형태소를 음절 단위 이상으로 앞 음절의 종성까지 분리하여 용언의 정확한 식별을 방해하는 예도 존재한다. '접고'라는 어절에서 '접-'이라는 용언을 정확히 분리해내지 못하고, '저-'를 용언으로 인식하게 되는 것이다. 반면, 음절 기반 분절은 '-은/는', '-이/가'와 같은 문법 요소들을 무조건적으로 분리하고, 단어 사전에 존재하는 음절들의 조합으로 모든 개체를 표현할 수 있으므로, 개체명 인식 및 의미역 결정 태스크에서 타 분절과 비교하여 유리한 성능을 보인다.

한편, 체언으로부터 조사를 분리하는 것이 핵심인 그룹 내 타 태스크와 달리, 의존 구문 분석 태스크의 경우 단어 분류 태스크임에도 불구하고, 서술어를 포함한 모든 문장 성분 간의 의미 관계 파악이 필요하므로, 문장 그룹과 유사한 경향을 보인다. 즉, 자모 단위 분절이 서술어의 통사적 변화를 포착하여, 가장 우수한 성능을 보여준다. 결국, 단어 분류 태스크에서는 음절 단위 분절이 체언과 문법 형태소의 분리에, 자모 단위 분절이 문장 성분 간 지배 관계 파악에 각각 우수한 성능을 보임을 알 수 있다.

4.3.2 문장 분류 태스크

성능 분석 Table 3의 실험 결과에 따르면, 자모 단위로 언어학적 분절을 진행한 모델이 세 태스크에서 모두 평균적으로 가장 좋은 성능을 보여준다. 특히, 문장 간 유사도를 평가하는 STS 태스크에서는 평균 정확도(Accuracy, %) 기준 점수가 79.52점으로 가장 높다. 이는 어절 단위의 분절을 적용한 모델보다도 약 1점가량 높고, 음절 단위의 분절을 적용한 모델에 비해서는 2점 이상의 성능 차이를 보여준다. 이에 비해 형태소 단위의 분절과 음절 단위의 분절이 적용된 모델은 자모 단위 분절은 물론 어절 단위 분절에 비해서도 다소 뛰어나지 못한 성능을 보인다. 그중 음절 단위 분절은 가장 저조한 성능을 보였는데, KorNLI 데이터셋에서는 어절 단위 분절과 자모 단위 분절 기법보다 성능이 평균 정확도 (Accuracy, %) 기준 2.3점 이상 낮고, KorSTS 데이터셋에서는 스피어맨 상관 계수(Spearman correlation) 기준 어절 단위 기법보다 1.4점 이상, 자모 단위 분절보다 2.4점 이상 낮음을 알 수 있다. 실험 결과를 통해 우리는 한국어 문장 분류 태스크에서 한국어 문장을 자모 단위로 분절하여 서브워드 알고리즘에 적용하는 것이 태스크와 관계없이 일관적으로 우수한 성능을 보여줌을 알 수 있다.

언어학적 분석 형태소 단위의 분절은 앞서 단어 분류 태스크에서 언급했던 것과 동일하게, 형태소 분석기의 오류가 성능 저하로 이어질 우려가 있다. '먹다'라는 동사가 '먹었다', '먹겠다', '먹었었다', '먹었고'와 같이 변화한다고 할 때, 자모 단위 분절의 경우 모든 자모를 분해 후 재조립 하는 방식을 통해 '먹어/게'와 같이 음절군을 '-ㅆ고/다', '-ㅆ었다'와 같은 토큰과 분리함으로써 의미군을 파악 가능하며, 또한 '-ㅆ다'과 같이 앞 음절의 종성 혹은 '먹ㅎ-'처럼 뒷 음절의 초성까지 고려하여 서술어의 시제 및 피동 여부까지 파악할 수 있다. 반면, 음절 단위 분절의 경우 용언을 어미로부터 분리하는 것에는 효과적이나, 이것이 오히려 서술어의 의미 파악에 방해가 될 수 있다. 예를 들어, 선어말어미 '-았/었'에 의해 시제만 달라질 뿐 행위 자체의 의미는 같은 동사 '하다'와 '했다'를 '하-'에 '-ㅆ'이 붙어 유사한 것으로 이해하지 못하고, '하'와 '했' 음절을 서로 상관없다고 파악하게 될 수 있다.

결국, 자모 단위의 분절 전략의 활용을 통해 한국어의 음소문자적 특징을 효과적으로 포착하여, 문장 분류 태스크에서 가장 우수한 성능을 기록함을 확인 가능하다.

5. 결론

본 연구에서는 한국어 분류 태스크의 성능 결과와 직결되는 입력 자질의 품질 제고 관점에서, 가장 효과적인 언어학적 분절 전략은 무엇이냐는 핵심 주제에 대답하고자 하였다. 이 과정에서 한국어 분류 태스크에 어려움을 야기하는 한국어의 특징 두 가지인 교착어적 특성과 음소문자의 특성을 언어학적 측면에서 정성적으로 제시하였으며, 세부 태스크의 유형을 단어 그룹과 문장 그룹으로 나누어, 각 그룹에서 이러한 특징들을 잘 반영하여 일관적으로 우수한 성능을 내는 분절 전략이 무엇인지 검증하였다. 이를 위해 보통 공백 단위로 분절하는 알파벳 언어와는 달리, 더 세밀한 분절 전략을 필요로 하는

한국어를 대상으로 어절, 형태소, 음절, 자모 네 가지 단위의 분절 전략을 적용한 모델들 간의 성능 비교를 진행하였다. 실험 결과에 따르면, 단어 그룹에서는 음절 단위의 분절 전략이 한국어의 교착어적 특성을 효과적으로 반영하여 뛰어난 성능을 보였으나, 단어 분류와 문장 분류의 중간적 성격을 띠는 의존 구문 분석 태스크의 경우, 자모 단위의 분절이 가장 우수한 성능을 보였다. 문장 그룹에서는 자모 단위의 분절 전략이 적용된 모델이 음절 혹은 형태소의 문맥상의 의미 변화를 포착해냄으로써 평균적으로 가장 높은 성능을 보인다. 결론적으로, 본 연구를 통해 개별 언어, 특히 한국어의 특징을 반영한 분절 전략의 효과성을 확인하였으며, 단어 분류와 문장 분류 태스크 간의 경향성의 차이에 대해 정량적인 수치와 더불어 언어학적으로 비교 검증하였다. 이를 통해 한국어 분류 태스크 연구 분야에, 입력 자질 구성의 관점에서 객관적 지표 및 언어학적 이해도 향상을 제공할 것으로 기대된다.

REFERENCES

[1] Y. Liu. et al. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692.*

[2] J. Devlin, M. W. Chang, K. Lee & K. Toutanova. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[3] R. Sennrich, B. Haddow & A. Birch. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909.*
DOI : 10.18653/v1/P16-1162

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

[5] T. Kudo. & J. Richardson. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226.*

[6] M. Kim., Y. Kim., Y. Lim. & E. N. Huh. (2019, July). Advanced subword segmentation and interdependent regularization mechanisms for korean language understanding. In 2019 *Third World Conference on Smart Trends in Systems Security and Sustainablity (WorldS4)* (pp. 221-227). London : UK.
DOI : 10.1109/WorldS4.2019.8903977

[7] O. Kwon, D. Kim, S. R. Lee, J. Choi & S. Lee. (2021, April). Handling Out-Of-Vocabulary Problem in Hangeul Word Embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 3213-3221).
DOI : 10.18653/v1/2021.eacl-main.280

[8] S. Park, J. Byun, S. Baek, Y. Cho & A. Oh. (2018, July). Subword-level word vector representations for Korean. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2429-2438).
DOI : 10.18653/v1/P18-1226

[9] S. Lee, H. Jang, Y. Baik, S. Park & H. Shin. (2020). Kr-bert: A small-scale korean-specific language model. *arXiv preprint arXiv:2008.03979.*
DOI : 10.5626/jok.2020.47.7.682

[10] A. Matteson, C. Lee, Y. Kim & H. S. Lim. (2018, August). Rich character-level information for Korean morphological analysis and part-of-speech tagging. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 2482-2492).

[11] S. Moon. & N. Okazaki. (2020, May). Jamo Pair Encoding: Subcharacter Representation-based Extreme Korean Vocabulary Compression for Efficient Subword Tokenization. *In Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 3490-3497). Marseille : France.

[12] D. B. Cho, H. Y. Lee & S. S. Kang. (2021). An Empirical Study of Korean Sentence Representation with Various Tokenizations. *Electronics, 10(7),* 845.
DOI : 10.3390/electronics10070845

[13] K. Park, J. Lee, S. Jang & D. Jung. (2020). An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks. *arXiv preprint arXiv:2010.02534.*

[14] T. Kudo. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959.*
DOI : 10.18653/v1/P18-1007

[15] E. F. Sang & F. De Meulder. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050.*

[16] S. Park. et al. (2021). KLUE: Korean Language Understanding Evaluation. *arXiv preprint arXiv:2105.09680.*

[17] J. Ham, Y. J. Choe, K. Park, I. Choi & H. Soh. (2020). Kornli and korsts: New benchmark datasets for korean natural language understanding. *arXiv preprint arXiv:2004.03289.* DOI : 10.18653/v1/2020.findings-emnlp.39

[18] M. Ott et al. (2019). fairseq: A fast, extensible toolkit for sequence modeling. arXiv preprint arXiv:1904.01038. DOI : 10.18653/v1/N19-4009

김 진 성(Jin-Sung Kim) [학생회원]

· 2019년 2월 : 고려대학교 서어서문학과(문학사) 및 LB&C 융합전공(언어, 뇌, 컴퓨터학사)
· 2019년 1월 ~ 2021년 1월 : LG Display System Engineer 및 KPMG Consultant
· 2021년 3월 ~ 현재 : 고려대학교 컴퓨터학과 석사과정
· 관심분야 : 자연어 처리, 정보 추출, 딥 러닝
· E-Mail : jin62304@korea.ac.kr

김 경 민(Gyeong-Min Kim) [학생회원]

· 2017년 8월 : 백석대학교 정보통신학부 정보보호학과(공학사)
· 2018년 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : 딥 러닝, 자연어 처리, 기계독해, 지식표현
· E-Mail : totoro4007@korea.ac.kr

손 준 영(Jun-Young Son) [학생회원]

· 2021년 8월 : 동국대학교 정보통신공학과(공학사)
· 2021년 9월 ~ 현재 : 고려대학교 컴퓨터학과 석박사통합과정
· 관심분야 : 자연어 처리, 정보 검색
· E-Mail : s0ny@korea.ac.kr

박 정 배(Jeongbae Park) [정회원]

· 2002년 2월 : 백석대학교 컴퓨터학과(공학사)
· 2014년 8월 : 고려대학교 컴퓨터교육학과(이학석사)
· 2020년 2월 : 고려대학교 컴퓨터학과(공학박사)
· 2020년 7월 ~ 현재 : 고려대학교 Human-inspired AI 연구소 연구교수
· 관심분야 : Natural Language Processing, Educational Data Mining, Social Network Analysis
· E-Mail : insmile@korea.ac.kr

임 희 석(Heuiseok Lim) [종신회원]

· 1992년 : 고려대학교 컴퓨터학과(이학학사)
· 1994년 : 고려대학교 컴퓨터학과(이학석사)
· 1997년 : 고려대학교 컴퓨터학과(이학박사)
· 2008년 ~ 현재 : 고려대학교 컴퓨터학과 교수
· 관심분야 : 자연어처리, 기계학습, 인공지능
· E-Mail : limhseok@korea.ac.kr