한국어 기계 독해를 위한 언어 모델의 효과적 토큰화 방법 탐구

이강욱○, 이해준, 김재원, 윤희원, 유원호

삼성전자 삼성리서치
{kw.brian.lee, haejun82.lee, j109.kim, huiwon.yun, wonho.ryu}@samsung.com

Exploration on Tokenization Method of Language Model for Korean Machine Reading Comprehension

Kangwook Lee○, Haejun Lee, Jaewon Kim, Huiwon Yun, Wonho Ryu
Samsung Electronics Samsung Research

요 약

토큰화는 입력 텍스트를 더 작은 단위의 텍스트로 분절하는 과정으로 주로 기계 학습 과정의 효율화를 위해 수행되는 전처리 작업이다. 현재까지 자연어 처리 분야 과업에 적용하기 위해 다양한 토큰화 방법이 제안되어 왔으나, 주로 텍스트를 효율적으로 분절하는데 초점을 맞춘 연구만이 이루어져 왔을 뿐, 한국어 데이터를 대상으로 최신 기계 학습 기법을 적용하고자 할 때 적합한 토큰화 방법이 무엇일지 탐구 해보기 위한 연구는 거의 이루어지지 않았다. 본 논문에서는 한국어 데이터를 대상으로 최신 기계 학습 기법인 전이 학습 기반의 자연어 처리 방법론을 적용하는데 있어 가장 적합한 토큰화 방법이 무엇인지 알아보기 위한 탐구 연구를 진행했다. 실험을 위해서는 대표적인 전이 학습 모형이면서 가장 좋은 성능을 보이고 있는 모형인 BERT를 이용했으며, 최종 성능 비교를 위해 토큰화 방법에 따라 성능이 크게 좌우되는 과업 중 하나인 기계 독해 과업을 채택했다. 비교 실험을 위한 토큰화 방법으로는 통상적으로 사용되는 음절, 어절, 형태소 단위뿐만 아니라 최근 각광을 받고 있는 토큰화 방식인 Byte Pair Encoding (BPE)를 채택했으며, 이와 더불어 새로운 토큰화 방법인 형태소 분절 단위 위에 BPE를 적용하는 혼합 토큰화 방법을 제안 한 뒤 성능 비교를 실시했다. 실험 결과, 어휘집 축소 효과 및 언어 모델의 퍼플렉시티 관점에서는 음절 단위 토큰화가 우수한 성능을 보였으나, 토큰 자체의 의미 내포 능력이 중요한 기계 독해 과업의 경우 형태소 단위의 토큰화가 우수한 성능을 보임을 확인할 수 있었다. 또한, BPE 토큰화가 종합적으로 우수한 성능을 보이는 가운데, 본 연구에서 새로이 제안한 형태소 분절과 BPE를 동시에 이용하는 혼합 토큰화 방법이 가장 우수한 성능을 보임을 확인할 수 있었다.

주제어: 자연어 처리, 전이 학습, 토큰화, 언어 모델, 기계 독해

1. 서론

현재 자연어 처리 분야에서는 딥러닝이 활발하게 사용되고 있다. 특히 트랜스포머 구조 기반의 BERT [1]가 소개 된 이후, 사전 학습(Pre-training) / 본 학습(Fine-tuning)의 두 단계로 이루어지는 전이 학습(Transfer Learning) 기반의 방법론이 보편화되었다. 전이 학습이란 한가지 문제를 해결하기 위해 습득한 지식을 다른 연관된 문제의 해결에 활용해보고자 제안된 기계 학습 방법론이다. 자연어 처리 분야에서 전이 학습은 크게 1) 대규모 말뭉치로부터 비지도 학습 방식으로 전반적 언어 지식을 습득하는 사전 학습 단계와 2) 축적된 지식을 토대로 목표 과업 해결을 위한 모델을 지도 학습 방식으로 학습 하는 본 학습, 총 두 가지 단계로 이루어진다. 이러한 전이 학습 기반의 방법론에서는 본 학습의 토대가 되는 사전 학습 언어 모델의 품질이 최종 성능에 큰 영향을 끼친다. 사전 학습 과정에서 품질 좋은 언어 모델을 얻기 위해서는 학습 전 전처리 단계(Preprocessing)가 매우 중요한데, 특히, 언어 모델의 입력 데이터로 사용되는 텍스트의 분절 방법이 중요하다.

텍스트의 분절은 토큰화(Tokenization)라 지칭하는데, 영어 등 주요 언어 대상으로 다양한 토큰화 방법이 제안되어왔으나, 한국어에서 그렇게 좋은 성능을 나타내지 못하고 있다. 이러한 결과를 초래하는 이유 중 하나는 한국어에 존재하는 교착어 특성을 토큰화 단계에서 고려하지 못했기 때문이다. 물론, 교착어 특성을 고려하기 위한 한국어 특화 토큰화 방법에 관한 연구도 진행되어 왔으나, 주로 토큰화 자체의 성능에 집중한 연구 [2]가 이루어져 왔고, 최종 과업 및 기계 학습에의 적용을 염두하고 성능 향상에 도움이 되는 토큰화 방법이 무엇일지 알아보는 탐구 연구는 거의 진행된 바 없었다.

본 연구에서는 전이 학습 기반의 자연어 처리 방법론을 염두 하여 한국어 언어 모델을 학습할 때 효과적인 토큰화 단위를 알아보기 위한 비교 연구를 수행한다. 비교를 위한 토큰화 단위로는 가장 기본적인 토큰화 단위인 1) 음절, 2) 어절, 3) 형태소와 더불어, 말뭉치 내 음절간 공기 통계를 이용하는 토큰화 방법인 4) Byte Pair Encoding (BPE) [3] 및 본 연구에서 새로이 제안하

는 토큰화 방법인 5) 형태소 단위로 분절 후 BPE를 적용한 혼합 토큰화 방법을 사용했다.

성능 비교 연구를 위한 최종 과업으로는 질의 응답의 한가지 분야인 기계 독해(Machine Reading Comprehension)를 선택했다. 기계 독해란 질의 응답(Question Answering)의 하위 분류 중 하나로, 질문과 그 질문에 대한 정답을 추론할 수 있는 지문이 주어졌을 때 정답을 도출해내는 과업이다. 기계 독해 과업의 해결을 위해서는 주어진 질문 및 지문에 대한 이해가 필수적이므로, 사전 학습으로 학습한 언어 이해 지식을 본 과업에 활용하고자 시도하는 전이 학습 기반의 자연어 처리 방법론을 적용하여 실험하기 적합하다. 실험을 위한 기본 모형으로는 현재 가장 보편적으로 사용되고 있으면서도 좋은 성능을 보이는 BERT를 사용했으며, 실험 데이터로는 공개 한국어 기계 독해 데이터셋인 KorQuAD v1.0 [4]을 사용했다.

비교 실험의 결과 분석 결과, 어휘집 크기 감소 효과 및 사전 학습 결과 얻어진 언어 모델의 성능은 음절 단위 토큰화가 효과적이었으나, 최종 과업인 한국어 기계 독해에서는 형태소 단위 토큰화가 효과적임을 확인할 수 있었다. 또한, 종합 성능을 고려할 때에는 BPE 토큰화 방법이 한국어에서도 좋은 성능을 보이는 가운데 본 연구에서 제안한 형태소 단위로 분절 후 BPE를 적용하는 혼합 토큰화 방법이 가장 우수한 성능을 보이는 것을 확인할 수 있었다.

2. 관련 연구

기계 학습에 있어 토큰화의 목표는 학습 모형이 효과적으로 동작할 수 있도록 돕는 데 있으므로, 어휘집의 크기를 줄이면서도 입력 텍스트가 내포하고 있었던 의미를 잘 보존될 수 있도록 텍스트를 분절하는 것이 중요하다. 예를 들어, 음절 단위 토큰화의 경우, 어휘집의 크기가 작아진다는 장점이 있지만 말뭉치 내 음절 토큰의 공기(co-occurrence) 사례가 너무 많아 연관 관계를 유추하기 어렵고, 그 결과 학습이 어려워진다는 단점이 있다. 반면, 어절 단위 토큰화의 경우, 토큰의 의미 내포 능력이 커져 자연어 처리 과업에 대한 학습이 상대적으로 용이해진다는 장점이 있지만, 동시에 토큰의 고유성 증가에 따라 어휘집 크기가 커지게 된다는 단점이 있다.

해외에서는 어휘집을 줄이면서도 최종 과업 성능을 높이기 위한 효과적인 토큰화 방법에 대한 연구가 지속적으로 이루어져 왔다. 그 중, 현재 가장 대중적으로 활용되고 있는 방법은 Byte Pair Encoding (BPE) [3]이다. BPE는 기계 번역 분야 연구에서 어휘집의 크기를 줄이기 위해 제안된 토큰화 방법론으로 등장 빈도가 높은 토큰 쌍을 하나의 새로운 토큰으로 묶어 나가는 방식으로 동작한다. 구체적인 동작 방법은 다음과 같다. 1) 말뭉치 내 모든 텍스트를 음절 단위 토큰으로 분절한 뒤 공기 통계를 산출한다. 2) 말뭉치 내에서 가장 등장 빈도가 높은 토큰 쌍을 묶어 새로운 토큰을 만들고 어휘집에 추가한다. 3) 2)의 새로운 토큰을 만드는 규칙을 결합 규칙으로 추가한다. 4) 미리 정의한 최대 결합 규칙 수 K에 도달할 때까지 2)~3)의 과정을 반복한다. BPE를 적용하는 경우, 최대 결합 규칙 수 K를 조절하는 방법으로 필요에 맞게 어휘집의 크기를 조절할 수 있다는 장점이 있다.

사전 학습한 언어 모델을 본 과업에 활용하는 전이 학습 기반의 자연어 처리 모형은 현재 자연어 처리 분야에서 널리 사용되고 있는데, 그 기본적인 개념은 ULMFiT [5]에서 최초로 제안 되었다. 그 후, 최신 신경망 구조인 트랜스포머를 활용한 GPT [6]와 BERT [1] 등 응용 모형이 제안되었다. 그 중 현재 가장 좋은 성능을 보이는 모형인 BERT는 트랜스포머 구조의 일부분인 멀티 헤드 어텐션 및 레지듀얼 커넥션을 이용해 구현된 인코딩 블록만으로 구성되어 있다. BERT의 경우, 1) 문장 내에 존재하는 토큰들을 임의로 마스킹 한 뒤 예측하게 하는 과업과 2) 입력으로 들어오는 두 개의 문장이 연속된 문장인지 아니면 임의로 조합된 문장인지 예측하는 두 가지 과업을 이용해 사전 학습이 이루어진다. 이 후, 본 학습에서는 모든 입력에 삽입되는 특이 용도 토큰인 [CLS]이나 각 토큰의 히든 값을 이용하는 분류 과업에 대해 학습이 이루어 진다.

본 연구에서는 최신 기술을 적극적으로 활용하기 위하여 전이 학습 기반의 자연어 처리 방법론을 실현하기 위한 기본 모형으로 BERT를 사용하고, BPE를 토큰화 방법 중 하나로 채택하여 비교 실험을 진행한다.

3. 한국어 토큰화 방법

현재까지 기계 학습의 효율 및 효과를 개선하기 위한 다양한 토큰화 방법들이 제안되어 왔으나, 이러한 방법들은 주로 영어와 같이 사용자가 많은 언어에서의 동작에 초점을 맞춰져 있었다. 하지만, 타 언어와 달리 한국어는 고유의 특성, 예를 들어 고유의 어순과 변형 및 교착어 특성을 가지므로 좋은 성능을 보인다고 알려져 있는 종래의 토큰화 방법들을 단순 적용하는 것은 효과적이지 않을 수 있다.

본 연구에서는 한국어 데이터를 대상으로 한 전이 학습 기반의 자연어 처리 방법론에서 토큰화 방법이 사전 학습 및 본 학습에 미치는 영향을 알아보기 위해 다양한 토큰화 방법을 비교군으로 설정하여 비교 실험을 진행한다. 비교 실험을 위해 채택한 토큰화 방법 및 그에 대한 설명은 다음과 같다.

음절: 글을 구성하는 가장 기본적인 단위인 음절 단위로 텍스트를 분절했다. 이 때, 입력 데이터 효율화를 추구하기 위해 공백은 음절 토큰으로 사용하지 않고 제외 했다.

어절: 글을 구성하는 또 다른 기본 단위인 어절은 띄어쓰기로 서로 구분된다. 본 연구에서는 띄어쓰기를 구분자로 이용, 텍스트를 분절하여 어절 단위로 사용 했다.

형태소: 형태소는 의미를 가지는 가장 작은 글의 단위이다. 분절을 위한 형태소 분석에는 공개 라이선스

형태소 분석기인 Mecab-Ko[1]를 활용했다.

BPE (K): 2장에서 언급했던 BPE 토큰화방법을 이용해 텍스트를 분절했다. 괄호 안에 있는 숫자는 BPE의 최대 결합 법칙 수 K로, 숫자가 커지면 커질수록 토큰의 다양성이 커진다. 또한, 어절 단위로 분절되는 부분과 BPE 토큰화로 인해 분절되는 부분을 구분하기 위해, BPE로 인해 분절되는 토큰에는 특수 접미사 @@을 삽입했다.

또한, 이와 더불어 형태소 단위로 텍스트를 분절한 후 BPE를 적용하는 혼합 토큰화 방법을 새로이 제안, 비교 대상 토큰화 방법의 하나로써 설정한 뒤 실험을 진행했다.

형태소→BPE (5만): 먼저 형태소 분석을 수행하여 텍스트를 분절한 뒤 BPE 토큰화 방법을 적용하여 텍스트를 분절했다. 실험에 사용한 BPE 토큰화의 최대 결합 규칙 수는 임의로 5만을 설정했다.

아래 표 1은 비교 실험에 사용한 다양한 토큰화 방법을 이용하여 텍스트를 분절한 결과물 예시를 보여준다.

표 1. 비교 실험에 사용한 토큰화 방법 적용 예시

입력 텍스트: 삼성로에 갔다.	
음절	[삼, 성, 로, 에, 갔, 다, .]
어절	[삼성로에, 갔다.]
형태소	[삼성로, 에, 갔다, .]
BPE	[삼성@@, 로@@, 에, 갔다.]
형태소 → BPE	[삼성@@, 로, 에, 갔다.]

토큰화 결과 예시를 보면 음절 단위 토큰에는 의미가 전혀 내포되어 있지 않은 반면, 형태소 단위 토큰에는 의미가 온전하게 내포되어 있음을 확인할 수 있다. 어절 단위 토큰화 결과는 상대적으로 두 가지 이상의 복합적인 의미를 내포하고 있음을 볼 수 있다. BPE 토큰화 결과는 부분적인 의미를 담고 있음을 확인할 수 있는데, 이는 BPE 토큰화의 통계 기반 동작 방식이 통계적으로 자주 함께 등장하는 음절/토큰들은 하나의 의미를 이룰 가능성이 높다는 점을 반영할 수 있기 때문이라고 해석된다.

이후 4장에서는 이러한 토큰화 방법의 특성이 전이 학습 기반의 자연어 처리 방법론의 성능에 어떤 영향을 미치는 지 알아보기 위해 토큰화 방법에 따라 사전 학습 언어 모델과 기계 독해 과업에 대한 본 학습을 수행한 뒤, 토큰화 방법에 따라 어떻게 성능 차이가 나는 지 비교 분석을 수행한다.

4. 비교 실험 및 분석

4.1 사전 학습

4.1.1 학습 환경

전이 학습 기반의 자연어 처리 방법론에 토큰화 방법이 미치는 영향을 분석하기 위해 우선 3.1장에서 소개한 다양한 토큰화 방법으로 분절한 텍스트에 대해 BERT의 기본(Base) 모형을 이용하여 사전 학습을 진행했다. 사전 학습 데이터로는 한국어 위키피디아 덤프 및 자체적으로 축적한 말뭉치를 사용했다. 사전 학습에 사용한 말뭉치의 통계는 표 2와 같다.

표 2. 사전 학습에 사용된 한국어 말뭉치 통계

용량	문장 수	단어 수
약 7.3 GB	5,832,488	67,533,870

사전 학습을 위해서는 BERT 논문에서 제안된 기본 모델의 구성 및 사전 학습 방법을 참고했다. 트랜스포머 레이어 12, 히든 차원 수 768, 셀프 어텐션 헤드 12개를 사용하는 기본 모형 설정을 차용하여 러닝 레이트 $1 * e^{-4}$, 배치 크기 384(GPU 8개 * 48 배치)로 백만 스텝의 마스크 토큰 예측 과업 및 문장의 연속성 판단 과업 기반의 학습을 진행했다. 단, 음절 단위 토큰화 경우, 타 토큰화 방식에 비해 입력 시퀀스의 길이가 짧아진다는 특성이 있어, 256 토큰의 최대 길이 / 최대 40개 토큰 추측을 적용한 타 토큰화 방법과 달리 512 토큰의 최대 길이 / 최대 80 토큰 추측을 적용하여 사전 학습을 진행했다.

4.1.2 사전 학습 언어 모델 성능

다양한 토큰화 방법을 적용해 사전 학습한 언어 모델의 성능을 비교하기 위해 어휘집 크기와 언어 모델의 퍼플렉시티 (Perplexity)를 측정했다. 퍼플렉시티란 확률 모델이 샘플을 얼마나 잘 예측할 수 있는 지에 대한 척도이다. 퍼플렉시티는 언어 모델의 품질을 측정하기 위한 척도로 활용되곤 하는데, 언어 모델에서 퍼플렉시티란 어떤 시점에서 후보로 예상 될 수 있는 토큰의 개수와 같다. 퍼플렉시티를 산출하기 위한 수식은 다음과 같다.

$$\text{Perplexity} = \mathrm{P}(\mathrm{w}_1, \ldots, \mathrm{w}_\mathrm{n})^{\frac{1}{n}} = e^{-\Sigma_x p(x) log q(x)} \quad (1)$$

이 때, n은 어휘집 크기를 나타내며, p(x)와 q(x)는 각각 샘플 분포와 예측 분포를 뜻한다. 토큰화 방법에 따른 어휘집 크기와 사전 학습 언어 모델의 퍼플렉시티는 다음 표 3과 같다. 단, 어절과 형태소 단위 토큰화의 경우 어휘집 크기가 지나치게 커서 통상적인 컴퓨팅 환경에서 학습이 어려웠기 때문에 빈도수를 기반으로 10만 개의 어휘만을 선별하여 사용했다.

표 3. 토큰화 방법별 어휘집 크기 및 사전 학습 언어 모델의 퍼플렉시티

[1] https://bitbucket.org/eunjeon/mecab-ko

	어휘 수	퍼플렉시티
음절	16,265	1.69
어절	100,000	13.02
형태소	100,000	3.54
BPE (3만)	51,971	4.21
BPE (4만)	61,971	4.80
BPE (5만)	71,962	4.68
BPE (6만)	81,956	5.32
BPE (7만)	91,941	5.90
형태소 → BPE (5만)	83,273	2.86

표 3에 나타난 결과는 크게 두 가지 관점에서 분석될 수 있다. 먼저 어휘집 크기의 축소 관점에서 보자면, 음절 단위의 토큰화가 가장 좋은 성능을 보이는 것을 볼 수 있다. 또한, BPE를 적용하는 경우 타 토큰화 방법에 비해 효과적으로 어휘집 크기를 줄일 수 있다는 것을 확인할 수 있다. 언어 모델의 퍼플렉시티 관점에서 보자면, 예측 대상 토큰의 수(즉, 어휘집 크기)가 상대적으로 적은 음절 단위 토큰화 방법과 토큰 자체가 의미를 분명하게 내포하고 있는 형태소 단위 토큰화 방법이 작은 퍼플렉시티를 보임을 확인할 수 있다. 불확실성에 대한 척도인 퍼플렉시티는 작을수록 좋으므로 음절 단위 토큰화 방법과 형태소 단위 토큰화 방법이 상대적으로 언어 모델 학습에 효과적이라고 해석할 수 있다. BPE 토큰화 방법의 경우에도 양호한 퍼플렉시티를 보이는 것을 확인할 수 있다. 주목할만한 점은 제안 방법인 형태소 분절 후 BPE 분절을 적용하는 혼합 토큰화 방법이 어휘집 크기 대비 우수한 퍼플렉시티를 보인다는 점이다. 이러한 결과는 제안 방법이 형태소 토큰화의 의미 내포 능력과 BPE 토큰화의 어휘집 축소 능력을 동시에 지니고 있기 때문이라고 해석될 수 있다.

4.2 기계 독해

4.2.1 학습 환경

전이 학습 기반 자연어 처리 방법론의 두번째 단계인 본 학습에 토큰화 방법이 미치는 영향을 분석하기 위해 3.2 장에서 사전 학습한 언어 모델을 토대로 한국어 기계 독해 과업에 대한 본 학습을 진행했다. 성능 평가를 위해서는 공개 한국어 기계 독해 데이터셋인 KorQuAD v1.0을 이용했다. KorQuAD v1.0은 대표적인 영어 기계 독해 데이터셋인 SQuAD v1.1 [7]의 제작 방법을 참고하여 만든 한국어 기계 독해 데이터셋으로, 양질의 한국어 위키피디아 문서를 수집한 뒤 문단을 추출하고, 크라우드 소싱(Crowd Sourcing)을 통해 해당 문단에 포함된 내용에 대해 물어볼 수 있는 질문과 답변을 작성케 하는 방법으로 제작한 데이터셋이다. 총 1,560개 위키피디아 문서에서 추출한 10,645 문단을 토대로 66,181개의 질의응답 쌍을 생성했으며, 이 중 60,407개의 질의응답 쌍은 학습 데이터셋으로, 나머지 5,774개의 질의응답 쌍은 검증 데이터셋으로 분리해서 제공하고 있다. 그림 1은 KorQuAD v1.0 데이터셋 내에 포함된 지문(Context) / 질의응답 쌍의 구조를 보여주고 있다.


```
{
  "qas": [
    {
      "answers": [
        {
          "text": "IP68",
          "answer_start": 63
        }
      ],
      "id": "6535617-3-1",
      "question": "갤럭시 S7 엣지의 방수 방진 등급은 무엇인가?"
    }
  ],
  "context": "그리고 갤럭시 S5에 탑재되었다가 갤럭시 S6/S6 엣지에서 도로 삭제되었던 방수 방진을 다시 지원한다. 등급은 IP68로, … "
}
```


그림 1. KorQuAD v1.0 데이터셋의 구조

본 연구에서는 본 학습 과업으로 한국어 기계 독해 성능을 평가하기 위해 KorQuAD v1.0 학습 데이터셋을 이용하여 본 학습을 진행 한 뒤 검증 데이터셋을 이용하여 성능을 측정했다. 본 학습을 위해서는 BERT 논문에서 제안된 기본 모델의 구성 및 본 학습 방법을 참고했다. 러닝 레이트 $3 * e^{-5}$, 배치 크기 384로 전체 학습 데이터를 2번 볼 수 있도록(2 epoch) 학습을 진행했다.

학습이 끝난 뒤에는 학습된 모델을 이용, 검증 데이터셋 상에서 정답 예측을 수행한 뒤 KorQuAD 순위 사이트[2]에서 제공하는 평가 스크립트 활용하여 성능을 측정했다.

4.2.2 한국어 기계 독해 성능

다양한 토큰화 방법 별로 본 학습을 수행한 한국어 기계 독해 모델의 성능을 비교하기 위해 Exact Match 및 F1 점수를 측정했다. 각 척도에 대한 구체적인 설명은 다음과 같다.

Exact Match: 실제 정답과 정확하게 일치하는 예측치의 비율

F1 점수: 실제 정답과 예측치를 이용해 음절 단위로 산출된 정밀도(Precision) 및 재현율 (Recall)의 조화 평균

예를 들어 정답이 “자연언어 처리”, 예측치가 “자연어 처리”일 경우, 정답과 예측치가 정확하게 일치하지 않으므로 Exact Match 점수는 0점이 되고, 음절 단위로 고려했을 때 예측치가 정답의 부분 집합이 되므로 정밀도 1, 재현율 6/7이 되어 F1 점수는 0.9231이 된다.

각 토큰화 방법을 적용하여 분절한 텍스트를 이용하여 본 학습한 모델의 KorQuAD v1.0 검증 데이터셋 상에서의

[2] https://korquad.github.io/

한국어 기계 독해 성능은 표 4와 같다.

표 4. KorQuAD v1.0 검증 데이터셋에서의 성능 비교

	Exact Match	F1 점수
음절	62.05	77.22
어절	13.04	59.40
형태소	83.86	92.37
BPE (3만)	25.03	82.50
BPE (4만)	20.09	81.69
BPE (5만)	20.06	82.07
BPE (6만)	19.97	82.06
BPE (7만)	20.02	81.97
형태소 → BPE (5만)	83.95	92.46

본 학습 성능을 평가 하기 위해 두 가지 척도 중 보다 실질적인 성능을 나타내는 F1 점수를 토대로 표 4에 나타난 결과를 분석해보자면, 본 학습인 한국어 기계 독해 과업에서는 음절 단위 토큰화 방법과 어절 단위 토큰화 방법이 저조한 성능을 보이는 가운데 형태소 단위 토큰화 방법과 BPE 토큰화 방법이 좋은 성능을 보임을 확인할 수 있다. 이는 토큰이 의미를 담지 못하는 음절 단위 토큰화나 너무 복합적인 의미를 담게 되는 어절 단위 토큰화의 경우에는 성능이 떨어지고, 근복적이면서도 분명한 수준의 의미를 내포하게 되는 형태소 단위 토큰화나 BPE 토큰화의 경우에는 성능이 올라가는 것을 보여준다. 이러한 결과를 토대로, 한국어 기계 독해 과업에는 토큰 단위에서의 의미 내포 능력이 성능에 큰 영향을 미치는 것을 알 수 있다.

또한, 결과 표를 통해 본 논문에서 새로이 제안한 형태소 단위로 분절 후 BPE 분절을 적용하는 혼합 토큰화 방법이 가장 좋은 성능을 보였다는 것을 확인할 수 있는데, 이는 제안 토큰화 방법이 사전 학습 단계와 본 학습 단계에서 모두 효과적으로 동작할 수 있었기 때문이라 해석될 수 있다. 즉, 제안 방법은 효과적인 어휘집 크기 감소 및 언어 모델의 낮은 퍼플렉시티를 보장하면서도 토큰 단위에서의 의미 보유 능력을 유지할 수 있었기 때문에 최종 과업인 기계 독해에서 가장 좋은 성능을 보일 수 있었던 것으로 판단된다.

또한 표 4의 결과에서 형태소 분석을 거치지 않은 거의 모든 토큰화 방법의 경우에서 F1 점수에 비해 Exact Match 점수가 매우 떨어지는 것을 볼 수 있는데, 이는 토큰화 과정에서 조사를 완벽하게 분리해내지 못했기 때문이다. 아래 표 5는 BPE(5만)에 대한 예측 결과 예제이다.

표 5. BPE(5만)에 대한 예측 결과 예제

ID	예측 결과	실제 정답
6511152-3-0	방수 방진을	방수 방진
6511152-3-1	IP68로,	IP68
6135555-4-0	갤럭시 S7 엣지가	갤럭시 S7 엣지

표 5의 결과를 보면, BPE 토큰화만을 적용했을 시 조사나 문장 부호를 올바르게 분리해내지 못한 텍스트를 정답으로 예측하고 있음을 확인할 수 있다. 이는 조사나 문장 부호가 통상적으로 명사구나 서술어와 함께 등장하는 경우가 많아 공기 통계가 높아지게 되고, 이로 인해 BPE와 같은 통계 기반의 토큰화 방법론으로는 분절이 어려워지기 때문이라고 사료된다. 즉, 교착어 특성이 있는 한국어의 경우, 통계 기반의 방법론만으로는 정교한 토큰화를 할 수 없다는 결론을 도출할 수 있다. 기계 독해 과업을 위해 통계 기반의 방법론을 활용하기 위해서는 최소한 조사나 문장 부호를 분리해주는 과정을 토큰화 전단계나 최종 결과값의 후처리 단계로 별도 적용이 필요할 것이다.

4장에서 이루어진 사전 학습과 본 학습에서 이루어진 비교 실험 결과를 토대로 종합적인 결론을 내리자면, 교착어 특성을 가지는 한국어에서는 형태소 단위 분절 이후 BPE 분절을 적용하는 혼합 토큰화 방법이 가장 효과적인 것을 확인할 수 있었다.

5. 결론

본 연구에서는 다양한 토큰화 방법이 전이 학습 기반 자연어 처리 방법론의 성능에 미치는 영향을 알아보기 위하여 비교 분석 실험을 수행했다. 실험 결과, 사전 학습을 통해 학습되는 언어 모델의 경우, 가장 기본적 텍스트 단위인 음절 단위 토큰화가 가장 좋은 어휘집 축소 및 퍼플렉시티 감소 성능을 보이는 가운데 BPE 토큰화가 어휘집 축소와 퍼플렉시티 감소에 탁월한 효과를 보임을 확인할 수 있었다. 반면, 본 학습인 한국어 기계 독해 과업에서는 토큰의 의미 보유 능력이 보장되는 형태소 단위 토큰화가 가장 좋은 성능을 보이는 가운데, 분절된 토큰이 어느 정도의 의미를 보유하게 되는 BPE 토큰화가 좋은 성능을 보임을 알 수 있었다.

또한, 비교 실험 결과를 토대로 본 연구에서 제안한 토큰화 방법인 형태소 분절 이후 BPE를 적용하는 혼합 토큰화 방법이 한국어 데이터를 대상으로 전이 학습 기반의 자연어 처리 방법론을 적용할 때 가장 효과적인 토큰화 방법이라는 결론을 내릴 수 있었다. 이는 제안 방법이 어휘집 크기를 축소하면서도 우수한 수준의 언어 모델 품질과 토큰의 의미 보유 능력을 보장할 수 있기 때문이라 분석 됐다.

향후에는 전이 학습의 개념을 교차 언어로 확장할 예정으로, 교차 언어에서 동작할 수 있는 범용 언어 모델을 구축하는 상황에서 효과적으로 활용될 수 있는 최적의 교차 언어 토큰화 방법을 탐구하기 위해 추가 연구를 진행할 예정이다.

참고문헌

[1] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, In Proceedings of NAACL-HLT, pp. 4171-4186, 2019.

[2] 이찬희, 이동엽, 허윤아, 양기수, 임희석, 음절 단위 및 자모 단위의 Byte Pair Encoding 비교 연구, 한글

및 한국어 정보처리 학술대회 논문집, pp. 291-295, 2018.
[3] R. Sennrich, B. Haddow and A. Birch, Neural Machine Translation of Rare Words with Subword Units, In Proceedings of ACL, pp. 1715-1725, 2016.
[4] 임승영, 김명지, 이주열, KorQuAD: 기계독해를 위한 한국어 질의응답 데이터셋, 한국정보과학회 한국소프트웨어종합학술대회 논문집, pp. 539-541, 2018.
[5] J. Howard and S. Ruder, Universal Language Model Fine-tuning for Text Classification, In Proceedings of ACL, pp. 328-339, 2018.
[6] A. Radford, K. Narasimhan, T. Salimans and I. Suskever, Improving Language Understanding by Generative Pre-Training, Technical Report of Open AI, 2018.
[7] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, SQuAD: 100,000+ Questions for Machine Comprehension of Text, In Proceedings of EMNLP, pp. 2383-2392, 2016.