한국어 형태소 언어 단위에 기반한 언어모델 구축 및 의존구문분석 모델 적용

2022 전기 졸업과제 착수보고서

부산대학교 전기컴퓨터공학부 정보컴퓨터전공

4조 A분과

노답삼형제

201724443 김준기

201724465 박기훈

201724568 정대성

목차

1. 과제 배경

구글에서 배포한 자연어 처리 언어 모델인 BERT(Bidirectional Encoder Representations from Transformers)는 transformer의 양방향 encoder로 문맥을 양방향으로 이해해서 숫자의 형태로 바꾸는 딥러닝 모델이라고 할 수 있다. BERT는 현재 전세계에서 가장 많이 쓰이는 자연어 처리 언어 모델이다. 하지만 내용어(명사, 동사 등)와 기능어(조사, 어미 등)가 공백으로 구분된 언어인 굴절어인 영어의 언어 단위 구성 방법이 적용되어 내용어와 기능어이 결합하여 어절을 구성하는 교착어인 한국어에는 적합한 방식이 아니다. 이에 ETRI(한국전자통신연구원)는 BERT를 기반으로 하지만 한국어에 최적화된 언어 모델인 KorBERT를 개발했고 이는 실제로 동일 학습데이터에서 BERT 대비 4.5% 우수한 성능(기계 독해에서는 4.3% 향상, 단락순위화에서는 7.4% 향상)을 보여주었다. 하지만 KorBERT도 100% 완벽하다고는 할 수 없어 개선의 여지가 남아있는 모습을 보여주고 있다. 이에 우리는 BERT가 아닌 RoBERTa, ELECTRA 등을 기반으로 하고 언어학적 처리 방식(은/는, 이/가, 을/를 등 의미와 성분이 같으나 두 가지 형태로 표현되는 조사 등)을 통일하는 식으로 하여 KorBERT보다 더 나은 성능을 보이는 한국어에 최적화된 자연어 처리 언어 모델을 개발할 것이다.

2. 용어 정리

2.1 트랜스포머

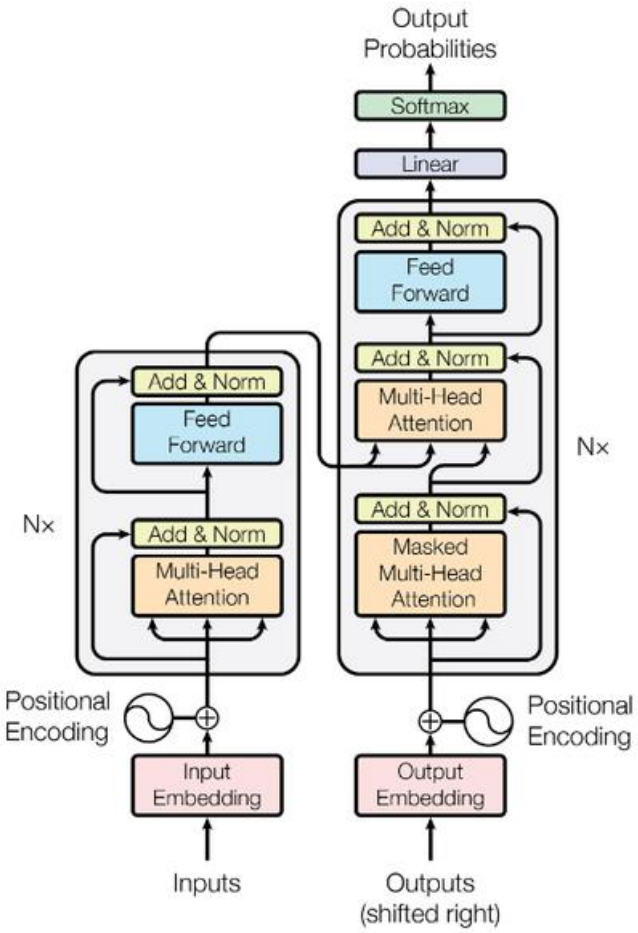

그림 1 트랜스포머의 구조(왼쪽이 인코더, 오른쪽이 디코더)

트랜스포머는 2017년 구글이 제안한 sequence-to-sequence 모델이다. 여기서 시퀀스란 단어(word)와 같은 무언가의 나열을 의미한다. sequence-to-sequence는 특정 속성을 지닌 시퀀스를 다른 속성의 sequence로 변환하는 작업을 말한다.

트랜스포머는 sequence-to-sequence 과제 수행에 특화된 모델로 인코더(encoder)와 디코더(decoder) 두 개의 파트로 구성된다. 인코더는 소스 시퀀스의 정보를 압축해 디코더로 보내주는 역할을 담당한다. 인코더가 소스 시퀀스를 압축하는 과정을 인코딩(encoding)이라고 한다. 그리고 디코더는 인코더로 받은 소스 시퀀스 정보를 받아서 타겟 시퀀스를 생성한다. 디코더가

타겟 시퀀스를 생성하는 과정을 디코딩이라고 한다. 인코더의 입력은 소스 시퀀스이고 디코더의 입력은 타겟 시퀀스의 일부이다.

이 트랜스포머를 기반으로 한 BERT 같은 경우는 사전학습 언어 모델이다. 사전학습 모델이란 학습 Label을 사람이 직접 만들지 않고 스스로 만들고 그 데이터를 학습함으로써 언어의 기본 소양을 쌓는 방식인 자기지도학습(Self-supervise Learning) 방식으로 만들어진 모델이다. BERT는 기본적으로 Wiki나 Book data와 같은 라벨링되지 않은(Unlabeled) 대용량 데이터로 모델을 미리 학습시킨 뒤, 특정 태스크(Task)를 가지고 있는 라벨링된 데이터로 전이학습(Transfer Learning)을 하는 사전학습 언어 모델이다.

단, 사전학습 모델은 모델 자체로 특정 기능을 수행할 수 없으므로 사전학습 모델을 기반으로 특정 태스크를 위해 한 번 더 학습하는 전이학습 즉, 파인 튜닝(Fine Tuning) 단계를 거쳐 딥러닝 모델을 미세하게 조정하는 학습 과정을 거쳐야만 여러 다양한 태스크에 활용이 가능해진다.

2.2 토크나이저

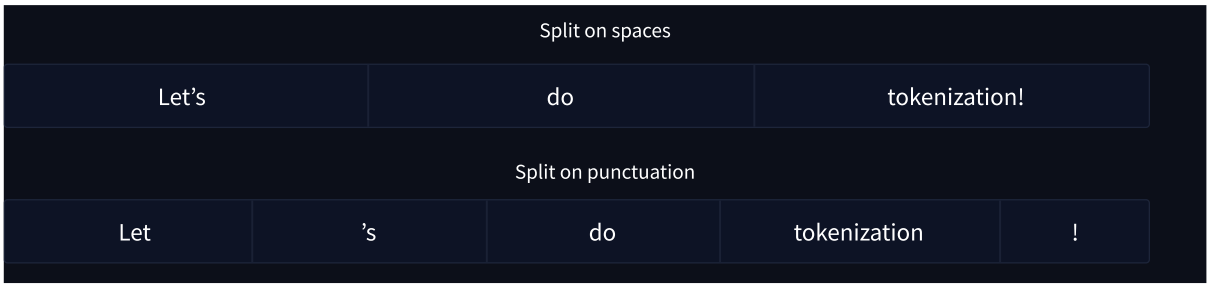

그림 2 단어 기반 토큰화

토크나이저는 NLP 파이프라인의 핵심 구성 요소 중 하나로 입력된 텍스트를 모델에서 처리할 수 있는 데이터로 변환하는 것이다. 가장 먼저 단어기반으로 토큰화 할 수 있다. 단어 기반 토크나이저로 특정 언어를 완전히 커버하기 위해서는 해당 언어의 모든 단어에 대한 식별자가 필요하고 이는 엄청나게 많은 토큰을 생성한다. 그리고 “dog”와 “dogs”가 유사한 단어임을 파악하기 어렵다.

그림 3 문자 기반 토큰화

다음으로 텍스트를 단어가 아닌 문자 기반으로 나누는 문자 기반 토큰화가 있다. 모든 단어들이 문자를 가지고 만들어질 수 있기 때문에, out-of-vocabulary 토큰이 훨씬 적다. 하지만 모델에서 처리해야할 토큰이 매우 많아진다는 점과 각 토큰이 문자 기반이므로 직관적으로 의미 파악이 어렵다는 한계점이 있어 위 두 가지 방식의 장점을 활용한 하위단어 토큰화 방식이 있다.

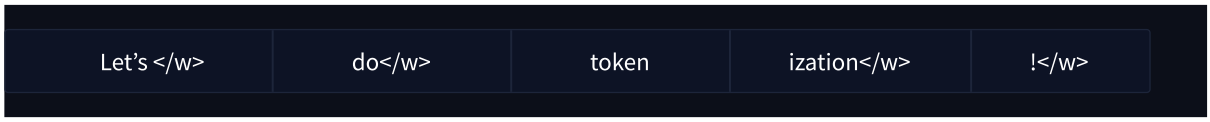

그림 4 하위단어 기반 토큰화

하위단어 토큰화는 빈번하게 사용하는 단어는 더 작은 하위단어로 분할하지 않고 희귀 단어를 의미 있는 단어로 분할한다. 예를 들어 tokenization이라는 단어는 희귀 단어로 간주되어 token과 ization으로 분할될 수 있다. 이 둘은 둘

다 독립적인 하위단어로 더 자주 출현할 가능성이 높으며 동시에 tokenization이 token과 ization의 합성의미로 유지된다.

2.3 의존구문분석

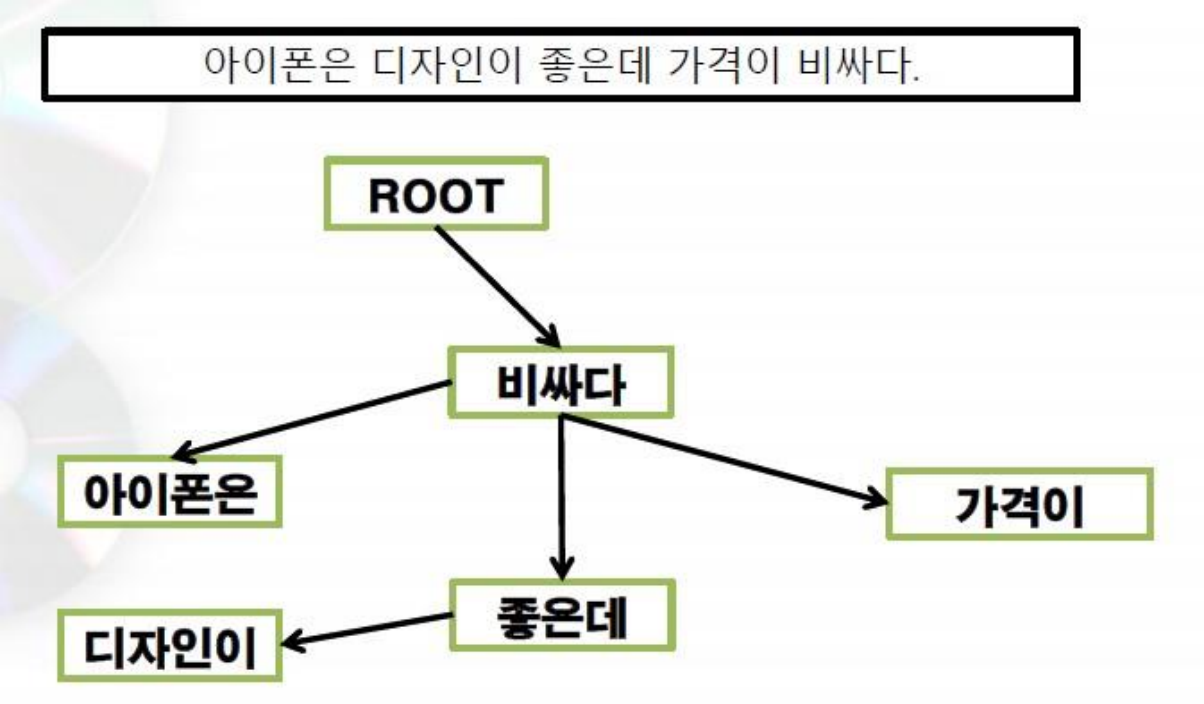

그림 5 의존구문분석의 예시

구문 분석은 문장의 구조를 이해하며 구조적 중의성을 해결하는 것이다. 일반적으로 한국어는 어순 배열의 자유도가 높고 문장 성분의 생략이 빈번한 특성이 있기 때문에 여러 구문 분석 방법 중 의존 구문 분석(dependency parsing)이 널리 사용되었다. 의존 구문 구조는 지배소(head)와 피지배소(modifier)로 구성되며 단어 간의 의존 관계로 표현된다.

3. 연구 방향

구글에서 배포한 자연어 처리 언어 모델인 BERT(Bidirectional Encoder Representations from Transformers)는 현재 세계적으로 가장 많이 쓰이고 있는 모델이다. 하지만 내용어(명사, 동사 등)와 기능어(조사, 어미 등)가 공백으로 구분된 언어인 굴절어와는 달리 내용어와 기능어가 결합되어 어절을 구성하는 교착어인 한국어에는 적합한 방식이 아니다. 이에 ETRI에서는 BERT를 기반으로 한국어에 최적화된 언어모델인 KorBERT를 개발했고 동일 학습데이터에서 BERT 대비 4.5% 성능 개선을 이루어 냈다.

	구글 배포 모델 언어 단위	**ETRI KORBERT 언어 단위**
접근 방법	영어의 언어 단위 구성 방법 적용 내용어와 기능어가 공백으로 구분된 언어(굴절어)	한국어 의미의 최소 단위인 형태소 기반 언어단위 내용어와 기능어가 어절로 결합된 언어(교착어)
언어 단위 구축 방법	말뭉치에서 통계적으로 추출한 음절 단위 n-gram 적용방법(BPE)	형태소 분석 이후, 형태소 단위에 대해서 BPE 적용 방법
언어 단위 적용 결과	한국 ##어 단 ##어는 형태 ##소로 구 ##성된 ##다.	한국어/NNp_ 단어/NNG_ 는/JX_ 형태소/NNG_ 로/JKB_ 구성/NNG_ 되/XSV_ s다/EF_ ./SF_
동일 학습데이터 평가결과	Google wordpiece 모델 - 기계독해: 90.68% - 단락순위화: 66.3%	KorBERT 형태소 모델 - 기계독해: 95.02% (+4.3%) - 단락순위화: 73.7% (+7.4%)

표 1 BERT와 KorBERT의 성능 비교(출처: ETRI)

ETRI에서 만든 KorBERT는 구글이 발표한 BERT 언어 모델을 기반으로 한국어의 특성을 고려하여 개선시킨 모델이다. 하지만 KorBERT는 현재로서는 최신형이 아니라고 볼 수 있는 BERT를 기반으로 개발되었다. BERT 언어 모델이 발표된 후로도 더 나은 언어 모델을 설계하기 위한 노력이 계속되었다. Undertrained된 BERT 모델을 개선한 RoBERTa와 GAN과 비슷한 구조로 학습의 효율을 향상시킨 ELECTRA 모델 등이 있다. 우리는 KorBERT와는 달리 RoBERTa, ELECTRA 등을 기반으로 모델을 설계하려 한다. 그리고 한국어의 조사중에 은/는, 이/가, 을/를 등 의미와 성분이 같으나 형태가 다른 조사들을 하나로 통일하여 처리하는 방식 등으로 기존의 KorBERT 보다 좀 더 개선된 모델을 개발하려고 한다.

4. 개발 일정 및 역할 분담

4.1 개발 일정

<table>
<tr><td colspan="3">5월</td><td colspan="4">6월</td><td colspan="5">7월</td><td colspan="4">8월</td><td colspan="4">9월</td></tr>
<tr><td>15</td><td>22</td><td>29</td><td>5</td><td>12</td><td>19</td><td>26</td><td>3</td><td>10</td><td>17</td><td>24</td><td>31</td><td>7</td><td>14</td><td>21</td><td>28</td><td>4</td><td>11</td><td>18</td><td>25</td></tr>
<tr><td colspan="3">딥러닝
스터디</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td colspan="4">기본 모델 작성</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="5">모델 최적화</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>중간
보고서
작성</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="6">예외 처리</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td colspan="3">최종 테스트</td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>최종
보고서
작성 및
발표</td></tr>
</table>

4.2 역할 분담

이름	역할
김준기	한국어에 특화된 토큰화 방법론 한국어의 단어를 자음과 모음위주로 분리한 후 서브 워드 단위 인코딩(wordpiece 맞추기) 교착어의 특성을 고려하여 형태소 분석기 framework를 사용한 데이터 전처리
박기훈	서브 워드 분절 알고리즘, Byte Pair Encoding, Unigram Language Model Tokenizer등 조사 영어와 한국어에서 다른 토큰화 방법에 대한 연구 및 알고리즘화 테스트 데이터셋 생성
정대성	머신러닝 학습을 위한 모델 개발 및 실제 학습(언어모델 실험 및 의존구문분석 모델 실험) 이를 이용한 기존의 영어 기반 subword 언어모델과 비교 평가 개선사항 분석 및 적용