세분화된 한국어 형태소에 기반한 사전학습모델 및 의존구문분석 모델 개발

2022 전기 졸업과제 중간보고서

부산대학교 전기컴퓨터공학부 정보컴퓨터전공

4조 A분과

노답삼형제

201724443 김준기

201724465 박기훈

201724568 정대성

지도교수: 권혁철

목차

1 요구조건 및 제약 사항 분석에 대한 수정사항

1.1 요구조건

한국어는 단어에 접사가 붙어야 그 뜻이 최종적으로 결정된다. 예를 들어, 나에 은/는/이/가 등이 붙으면 주격이 되고 을/를 등이 붙으면 목적어가 된다. Wordpiece 기반의 사전학습 언어모델인 BERT는 형태소 단위가 아니라 조사 분류를 안 해서 구조적인 정보를 반영하지 않아 한국어에는 상대적으로 알맞은 언어모델이 아니다.

이에 이를 극복하기 위한 ETRI가 개발한 KOBERT가 존재하나 TTAS 표준을 따르다 보니 상대적으로 형태소 분류 세분화가 KLTagger에 비해 부족하다. 예를 들어 자연어 처리라는 단어가 있다면, KLTagger는 자연어를 일반명사, 처리를 동작성명사로 구분하는 반면, TTAS 표준은 자연어와 처리를 모두 NNG라는 하나의 태그로 분류한다.

그렇기에 우리는 구조적으로 더 세분화된 분석을 하는 KLTagger와 POS 임베딩이 가능한 사전학습 언어모델을 개발하려고 한다.

1.2 기존 제약사항 및 수정사항

기존에는 POS 임베딩만을 추가하여 사전학습 언어모델을 개선하려고 했으나, 현재는 KLTagger의 결과를 TTAS로 매핑하지 않고 POS 임베딩과 Vocab 생성 방식을 활용한 모델 2가지, TTAS로 매핑하고 POS 임베딩과 Vocab 생성 방식을 활용한 모델 2가지, TTAS 표준 형태소 기반으로 규칙을 적용하고 있는 파서 모델을 개선한 모델을 1가지, 총 5가지 모델을 개발하는 것으로 목표 수정이 이루어졌다. 다른 언어 모델이 TTAS 표준임에도 불구하고 매핑하지 않고 KLTagger만의 결과를 활용하는 모델을 개발하는 이유는 KLTagger와 TTAS 표준의 형태소 분석 개수가 KLTagger는 53개, TTAS는 41개로 달라 KLTagger 기반시 구조적으로 더 세분화된 분석 가능해 성능적으로 더 나은 면모를 보일 것으로 예측이 되기 때문이다.

2 설계 상세화 및 변경 내역

2.1 BERT의 구조

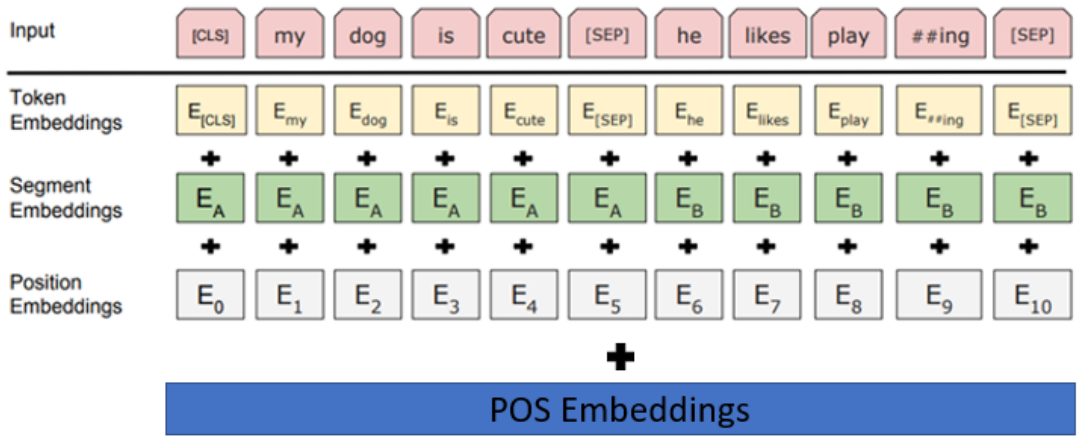

그림 1 BERT의 Input 구조 및 POS Embeddings

BERT는 Input으로 받은 문장을 3가지로 Embedding한다. 문장을 Subword 단위로 나누어 Token으로 구분하는 Token Embeddings, 문장을 구분하는 Segment Embeddings, Token의 순서를 지정하는 Position Embeddings 등이 있다. 단, 이 세 가지 Embeddings만으로는 구조적 정보를 분석할 수가 없어 한국어 분석에는 어려움을 겪는다. 이에, POS(Part of Speech: 형태소) Embeddings를 추가함으로써 한국어 분석에 도움을 주는 식으로 Parser 모델을 개선할 방식이다.

2.2 Parser 모델

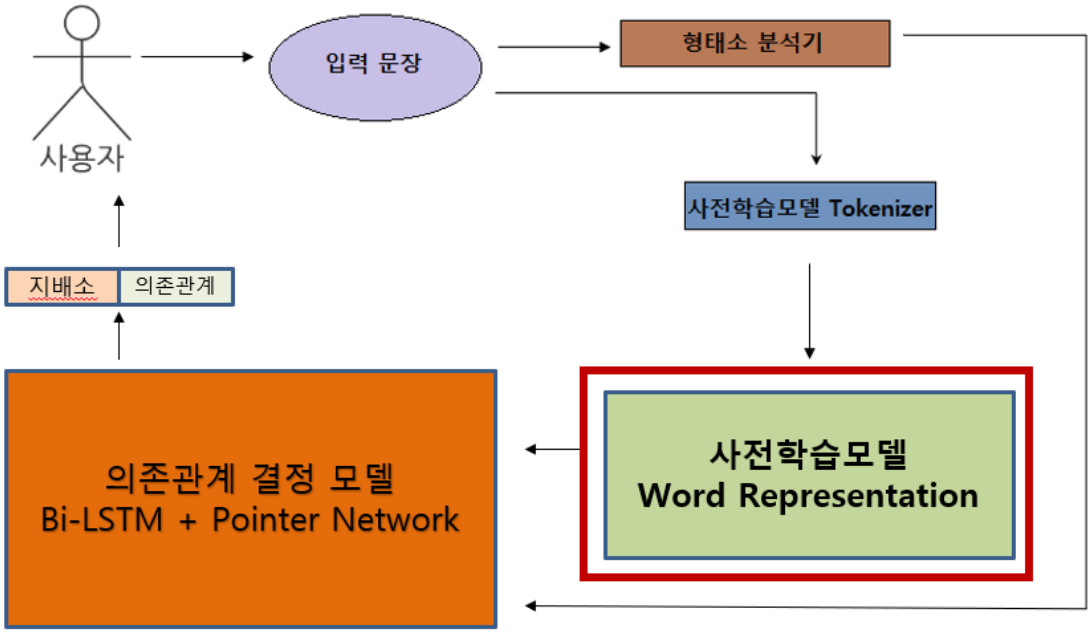

그림 2 PARSER 모델의 구조

PARSER 모델의 구조는 위와 같다. 사용자가 문장을 입력하면 토크나이저와 형태소 분석기로 각각 Input으로 들어가게 된다. 사전학습모델 Tokenizer가 처리한 결과는 사전학습모델의 Input으로 들어가게 되고, 사전학습모델과 형태소 분석기가 처리한 결과가 의존관계 결정 모델에 Input으로 들어가서 최종적으로 입력 문장의 지배소와 의존관계를 출력하여 사용자에게 알려주는 구조로 이루어져 있다.

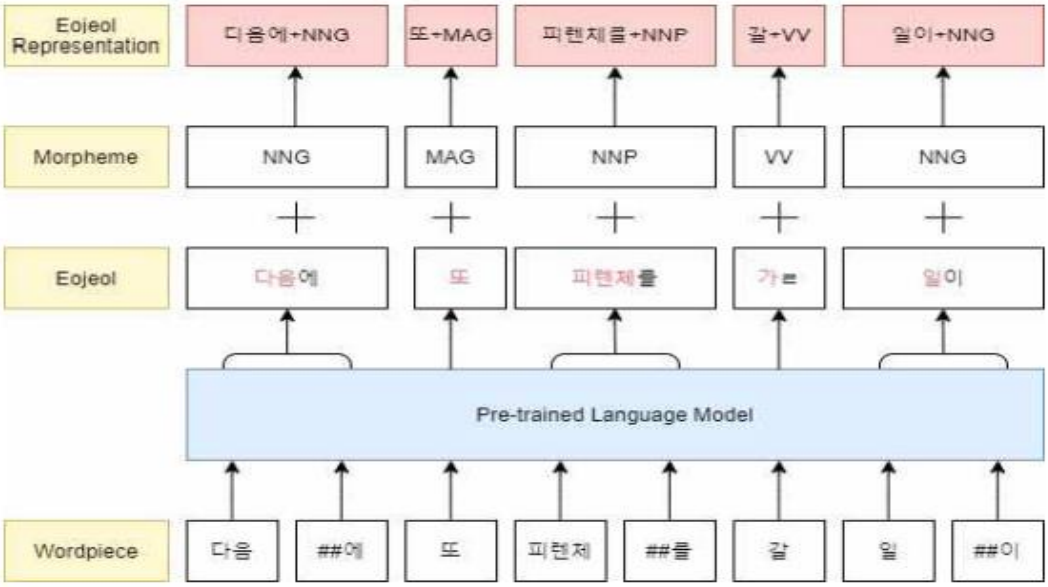

그림 3 파서 모델의 구조 일부 예시 1

위 그림은 파서 모델의 구조를 일부를 예시와 함께 나타낸 그림이다. '다음에 또 피렌체를 갈 일이'의 문장을 입력하면 해당 문장을 Wordpiece 단위로 Tokenizing 후, 사전학습 언어모델로 분석하면 '다음에, 또, 피렌체를, 가ㄹ, 일이'라는 어절들과 각 어절에 해당하는 형태소 특성인 NNG, MAG, NNP, VV, NNG를 매칭시켜 어절 표상을 생성한다.

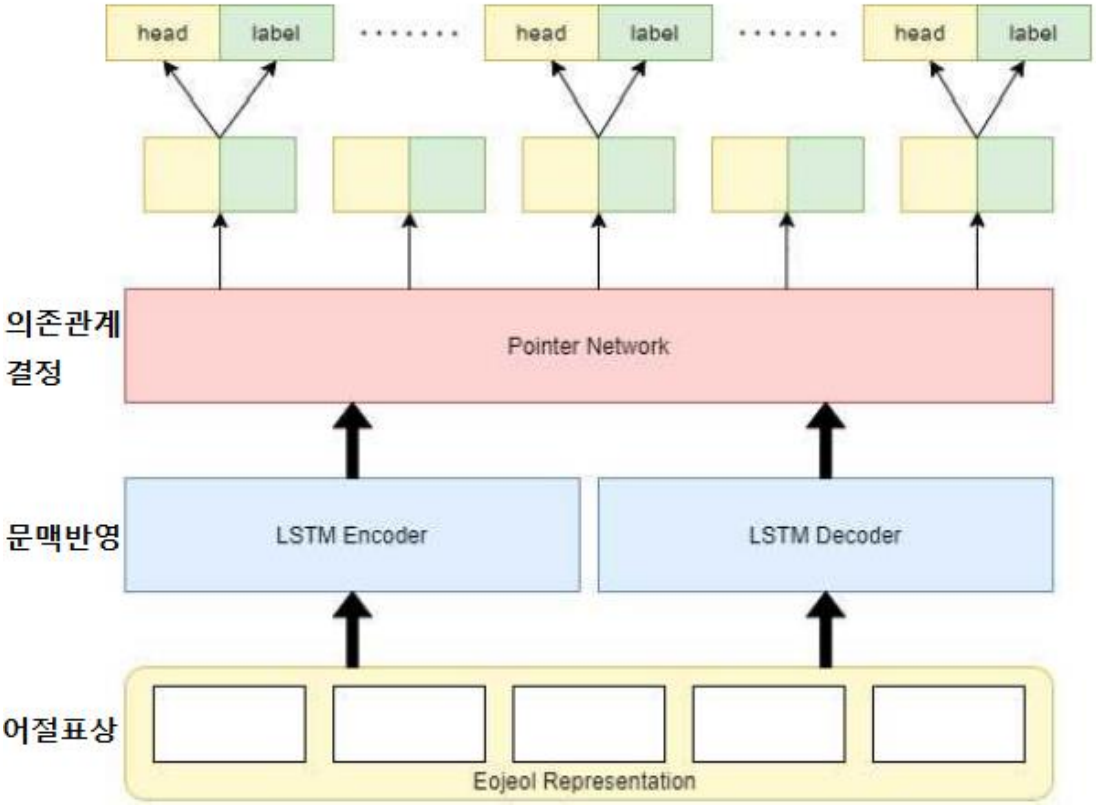

그림 4 파서 모델의 구조 일부 예시 2

이후, 이 어절 표상이 가변 길이 입력 순서와 가변 길이 출력을 지원하는 Long Short Term Memory Encoder와 Decoder로 들어가 입력한 어절 표상의 문맥을 반영하여 출력하게 된다. 마찬가지로 가변 길이 입출력을 지원하는 Pointer Network로 각 어절 표상의 의존관계를 결정해서 Head와 Label부로 나누어 결과를 출력한다. Head는 문장내에서 해당 어절이 의존하는 어절의 위치에 관한 인덱스를 나타내며, Label은 피지배소와 지배소의 연관관계를 나타낸다.

3 갱신된 과제 추진 계획

5월			6월				7월					8월				9월			
15	22	29	5	12	19	26	3	10	17	24	31	7	14	21	28	4	11	18	25
딥러닝 스터디																			
			데이터 수집																
							데이터 전처리												
							품사 태깅												
										중간 보고서 작성									
												TTAS 기반 사전학습모델 개발							
												KLTagger 기반 사전학습모델 개발							
															모델 학습 및 최적화				
															예외 처리				
																		최종 보고서 작성 및 발표	

4 구성원 별 진척도

이름	역할 분담
김준기	데이터 수집, 중간 보고서 작성
박기훈	데이터 전처리
정대성	품사 태깅

5 보고 시점까지의 과제 수행 내용 및 중간 결과

5.1 데이터 추출

우선 데이터 전처리 및 형태소 태깅에 사용할 데이터를 추출했다. 모두의 말뭉치에서 19GB의 문장들, 2018년도 동아일보 기사에서 랜덤으로 10GB의 문장들을 추출해 총 29GB의 문장들을 데이터셋으로 활용하기로 했다.

5.2 형태소 태깅

형태소 분석기(Tagger: 원시말뭉치를 형태소 단위로 쪼개고 각 형태소에 품사 정보를 부착하는 작업을 수행하는 프로그램)를 사용해 문장에서 형태소 분석을 진행했고, 이후 해당 결과를 저장했다. 예를 들어, '나는 자연어 처리를 공부하는 학생이다' 라는 문장이 있다면, 이를 KLTagger를 이용해 분석하면 '나(인칭대명사) / 는(보조사) / 자연어(일반명사) / 처리(동작성명사) / 를(ETM) / 공부(일반명사) / 하다(동사) / 는(관형형전성어미) / 학생(일반명사) / 이다.' 라는 결과로 분석되어 출력된다.

5.3 KLTagger -> TTAS로 매핑

현재 KLTagger에 Input Data를 넣어 나온 결과 값의 형태소를 한국어(KLTagger)에서 영어(TTAS)로 바꾸는 작업을 진행 중이다. TTAS란 Telecommunications Technology Association의 약자로, 정보통신 단체표준을 의미하는 단어이다. 현재 형태소 데이터셋의 표준, 사전학습 모델들의 형태소 정보도 TTAS 표준으로 사용하기 때문에 KLTagger의 출력 결과를 TTAS로 매핑하는 과정을 진행 중이다. 해당 과정은 아래와 같은 표에 맞춰서 진행했다.

고유명사	NNP	주격보격조사	JKS	보조용언	VX	명사형전성어미	ETN	물음표	SF
일반명사	NNG	형용사	VA	일반관형사	MMD	접속부사	ETN	관형사화접미사	XSN
수관형사	MMN	일반의존명사	NNB	화폐단위	NNB	일반부사	MAG	쌍점	SP
도량형단위	SW	동사	VV	여는큰따옴표	SS	형용사화접미사	XSA	명사	NNG
수접미사	XSN	종결어미	EF	닫는큰따옴표	SS	반점	SP	서수사	NR
단위성의존명사	NNB	온점	SF	한자	SH	상태성명사	NNG	느낌표	SF
줄임표	SE	접속조사	JC	여는괄호	SS	지시대명사	NP	분석불능	NA
타동사	VV	외국어	SL	닫는괄호	SS	기타기호	SW	반쌍점	SP
연결어미	EC	양수사	NR	자타동사	VV	인칭대명사	NP	기타문자	SW
동작성명사	NNG	이음표	SO	여는작은따옴표	SS	중간방점	SP		
일반접미사	XSN	목적격조사	JKO	닫는작은따옴표	SS	빗금	SP		
부사격조사	JKB	보조사	JX	인용격조사	JKQ	일반접두사	XPN		
동사화접미사	XSV	지정사	VCP	복수접미사	XSN	수접두사	XPN		
관형형전성어미	ETM	부정지정사	VCN	인용형어미	JKQ	감탄사	IC		
관형격조사	JKG	선어말어미	EP	자동사	VV	호격조사	JKV		

그림 5 KLTagger -> TTAS 매핑 표

5.4 개선할 모델 선정 결과

개선할 모델을 선정하기 위해 시중에 배포된 사전학습모델을 Train Set, Test Set은 각각 KLUE로 고정하고 성능을 테스트해보았다. 점수 산출은 UAS(Unlabeled Attachment Score: 의미론적 관계를 고려하지 않고 평가함), LAS(Labeled Attachment Score: 의미론적 관계를 고려하여 평가함)을 사용하였다. Test한 모델은 총 4개로 koBERT-base, Klue/RoBerta-base, Klue/RoBerta-Large, KoElectra-base 순으로 진행하였다. 결과는 다음과 같았다.

Train set	Test set	사전학습모델	UAS	LAS
KLUE	KLUE	koBERT-base	92.93	90.30
KLUE	KLUE	Klue/RoBerta-base	93.04	90.41
KLUE	KLUE	Klue/RoBerta-Large	93.48	90.57
KLUE	KLUE	KoElectra-base	**93.55**	**90.82**

KLUE의 Train set의 크기는 6000문장, Test set의 크기는 1000문장이다.

그림 6 사전학습모델별 성능 평가표

결과는 다음과 같이 나왔는데, KoElectra-base가 UAS, LAS 모든 부분에서 다른 사전학습 모델들 보다 뛰어난 성능을 보여 이를 개선할 모델로 선택하게 되었다.