KRBERT 임베딩 층에 따른 의미역 결정

서혜진[1], 박명관[1], 김유희[2]
동국대학교[1], 신한대학교[2]
seohj0951@gmail.com, korgen2002@naver.com, euhkim@gmail.com

Layerwise Semantic Role Labeling in KRBERT

Seo, Hye-Jin, Park[1], Myung-Kwan[1], Kim, Euhee[2]
Dongguk University[1], Shinhan University[2]

요 약

의미역 결정은 문장 속에서 서술어와 그 논항의 관계를 파악하며, '누가, 무엇을, 어떻게, 왜' 등과 같은 의미역 관계를 찾아내는 자연어 처리 기법이다. 최근 수행되고 있는 의미역 결정 연구는 주로 말뭉치를 활용하여 딥러닝 학습을 하는 방식으로 연구가 이루어지고 있다. 최근 구글에서 개발한 사전 훈련된 Bidirectional Encoder Representations from Transformers (BERT) 모델이 다양한 자연어 처리 분야에서 상당히 높은 성능을 보이고 있다. 본 논문에서는 한국어 의미역 결정 성능 향상을 위해 한국어의 언어적 특징을 고려하며 사전 학습된 SNU KR-BERT를 사용하면서 한국어 의미역 결정 모델의 성능을 살펴보였다. 또한, 본 논문에서는 BERT 모델에서 과연 어떤 히든 레이어(hidden layer)에서 한국어 의미역 결정을 더 잘 수행하는지 알아보고자 하였다. 실험 결과 마지막 히든 레이어 임베딩을 활용하였을 때, 언어 모델의 성능은 66.4% 였다. 히든 레이어 별 언어 모델 성능을 비교한 결과, 마지막 4개의 히든 레이어를 이었을 때(concatenated), 언어 모델의 성능은 67.9% 이였으며, 11번째 히든 레이어를 사용했을 때는 68.1% 이였다. 즉, 마지막 히든 레이어를 선택했을 때보다 더 성능이 좋았다는 것을 알 수 있었다. 하지만 각 언어 모델 별 히트맵을 그려보았을 때는 마지막 히든 레이어 임베딩을 활용한 언어 모델이 더 정확히 의미역 판단을 한다는 것을 알 수 있었다.

주제어: 한국어 의미역 결정, BERT, BI-LSTM CRF

1. 서론

인공지능(Artificial Intelligence; AI) 기술이 나날이 발전해가면서 이를 자연어 처리(Natural Language Processing; NLP)에 접목시키는 연구들이 많이 진행되고 있다. 자연어 처리는 컴퓨터가 사람처럼 언어를 이해할 수 있게 하는 기술로써, 대량의 텍스트에 함축되어 있는 의미를 파악하고 텍스트에 포함된 정보를 추출하거나 분류하는 기술을 일컫는다. 본 연구에서는 자연어 처리의 한 분야인 의미역 결정(Semantic Role Labeling)에 대해서 알아보고자 한다. 의미역(semantic role)은 서술어에 의한 행동이나 상태에 의해서 명사구가 갖는 의미 역할을 말하며, 의미역이 부여된 각 명사구는 논항(argument)이라고 한다[1]. 그림 1과 같이 의미역 결정은 문장 속 서술어와 (각 서술어에 속하는) 논항들 사이에서의 '누가, 무엇을, 어떻게, 왜' 등의 의미관계를 파악하는 것이다. 이러한 연구는 기계번역 및 질의응답, 정보추출과 같은 다양한 자연어 처리 연구에 응용되어 활용될 수 있다[2].

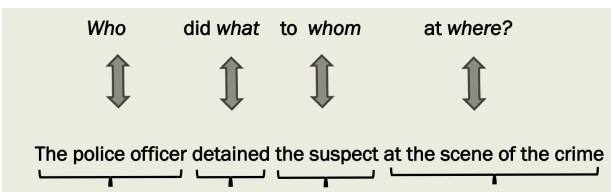


그림 1. 의미역 결정 과제 예시

비교적 최근에 구글이 공개한 인공지능 언어 모델인 Bidirectional Encoder Representations from Transformer (BERT)[3]는 다양한 자연어 처리 과제에서 상당히 높은 성능을 보였다. BERT는 Transformer 기반의 사전 훈련된 모델로써 Multi-Head Attention과 N개의 인코딩 블록(encoding block)으로 구성되어 있다. 본 논문에서는 SNU KR-BERT 모델을 활용하면서 한국어 의미역 결정을 살펴보고자 한다. 서울대에서 개발한 KR-BERT는 다른 한국어 BERT 모델들보다 한국어의 언어적 특징을 고려하면서 훈련되었다[4]. 또한 순차적 데이터(sequential data) 처리에 가까운 의미역 결정 모델의 성능을 높이기 위해서, BERT 모델에 Bidirectional Long short-term momory(Bi-LSTM)와 Conditional Random Field (CRF) 레이어(layer)를 추가하였다. 이와 같이 추가적으로 레이어를 쌓음으로써, 의미역 태그 사이의 의존성 전이 확률까지도 고려되기 때문에 의미역 결정 모델의 성능을 높일 수 있다[1]. 본 논문에서는 BERT Bi-LSTM CRF 모델의 성능 뿐만 아니라, BERT 모델 12개 인코딩 블록의 레이어 별 성능도 살펴보고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 BERT Bi-LSTM CRF 모델에 대해서 설명하고, 4장에서는 딥러닝 언어 모델 레이어 별 분석에 대해서 다루고, 5장에서는 실험 및 결과를 살펴본다. 마지막으로 6장에서는 논의점에 대해 기술한다.

2. 관련 연구

의미역 결정 연구는 격틀사전을 활용하는 방법과 말뭉치를 활용하는 방법으로 나뉜다. 격틀사전을 활용하는 방법은 서술어와 각 논항의 의미역들이 정리되어 있는 격틀사전을 이용하여 문장의 의미역을

결정하는 방법이다. 해당 방법은 정확도가 상당히 높지만 격틀사전을 구축하는데 많은 자원을 필요로 하고, 격틀사전에 구축되어있지 않은 의미역들을 결정하는데 어려움이 있다. 반면에, 말뭉치를 활용하는 방법은 의미역이 태깅된 말뭉치를 사용하여 기계학습으로 의미역을 결정하는 방법이다. 해당 방법은 말뭉치에 포함되어 있지 않은 의미역 결정에도 상당히 높은 정확도를 보이지만, 의미역이 태깅된 말뭉치를 구축하는데 어려움이 있다[1].

딥러닝 기술 발전에 따라서 최근 의미역 결정 연구에서는 말뭉치를 활용한 의미역 결정 언어 모델 연구가 활발히 진행되고 있다 [1,2,5-12]. 일반적으로 딥러닝을 활용한 언어모델은 품사태깅(part-of-speech tagging)[5]이나 구문 트리(syntactic tree) [6,7]와 같은 언어적 자질을 고려하면서 학습되었다. 예를 들면, [6]에서는 통사적 자질은 의미역 결정 언어 모델의 성능을 높이는데 필요하다고 주장하였다. 하지만 최근 의미역 결정 연구에서는 세부적인 언어적 자질 없이 훈련된 BERT 모델도 성능이 상당히 좋았다고 주장하였다. 구체적으로, [8]에서는 BERT 모델을 활용하여 영어 의미역 결정 언어 모델을 연구하였으며, 어휘나 통사적 자질을 부여하지 않고 BERT 모델을 학습 시켜도 언어 모델이 의미역 결정을 상당히 잘한다고 주장하였다. [9]에서는 한국어로 사전 학습된 ETRI KorBERT를 활용하여 한국어에서의 의미역 결정 언어 모델을 연구하였으며, Bi-LSTM CRF를 사용한 의미역 결정 언어 모델[10,11]보다 BERT Bi-LSTM CRF 언어 모델의 성능이 좋다고 주장하였다.

3. BERT Bi-LSTM CRF를 사용한 한국어 의미역 결정 언어 모델

그림 2와 같이 BERT Bi-LSTM CRF 모델은 출력 레이블을 결정하기 위해 BERT의 출력을 순차적 데이터 모델링에 적합한 Bi-LSTM 모델에 보냄으로써 출력 레이블 간의 의존성을 추가하고 CRF 레이어를 통해서 인접한 레이블의 의미역 정보를 참고하면서 최종 출력 레이블을 결정한다.

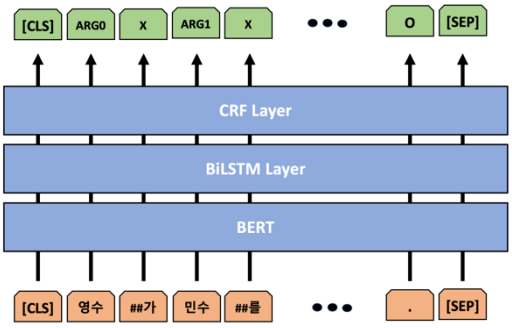


그림 2. BERT Bi-LSTM CRF 언어 모델[1]

4. 딥러닝 언어 모델 레이어 별 분석

딥러닝 모델이 다양한 자연어 처리 연구에서 높은 성능을 보이고 있지만, 실제적 과제(real-world tasks)의 적용엔 한계가 있다. 실제로 딥러닝 모델이 어떻게 작동하기에 다양한 자연어 처리 연구에서 좋은 성능을 이룰 수 있는지에 대해선 알려져 있지 않기 때문에 블랙박스(black box) 모델이라고도 한다[13]. 실제적 과제 적용의 한계를 극복하기 위해서는 블랙박스 모델을 좀 더 탐험적으로 연구하는 과정이 필요하다. 최근 들어, 히든 레이어(hidden layer)를 살펴보는 연구들이 늘어나고 있다[12,13,14,15]. [14]에서는 질의응답 언어 모델이 히든 레이어 별로 어떻게 달라지는지 살펴보았다. 그 결과, 초반 히든 레이어에서는 단어 별 의미 군집화(semantic clustering)를 형성하지 못하지만 중간 히든 레이어부터 정확하진 않지만 주제에 맞춰서 의미 군집화를 형성해 나간다고 한다. 이러한 과정을 거치면서 마지막 레이어에선 질문에 관련된 적절한 응답을 도출해 나갈 수 있게 된다고 주장하였다.

5. 실험 및 결과

본 논문에서는 히든 레이어 별 성능(performance)을 살펴보면서, 어떤 히든 레이어가 의미역 결정을 가장 잘 하는지에 대해서 알아보고자 한다. 구체적으로, 언어 모델의 성능을 크게 두 가지 방식으로 비교·분석 하였다. 우선 일반적으로 사용하는 마지막 히든 레이어의 언어 모델 성능을 살펴보았다. 그리고 히든 레이어에 따라 달라지는 언어 모델 성능을 살펴보았다.

의미역 결정 연구를 위해서 [9]에서 사용한 BERT Bi-LSTM CRF 언어 모델을 토대로 연구를 진행하였다. 표 1에서 살펴볼 수 있듯이 선행 논문과의 차이점은 다음과 같다. [9]와 본 연구의 가장 큰 차이점은 ETRI KorBERT 대신에 SNU KR-BERT를 사용하였으며, Korean prop-bank 대신에 NAVER와 창원대가 개최한 한국어 자연어처리 기술 대회에서 제공한 의미역 결정 데이터를 활용하였다는 점이다[2]. 표 2를 보면 사전 훈련 모델 구축

[1] 의미역 결정연구에서 ARG0는 주어, ARG1은 목적어, O는 의미역에 해당하지 않는 것, X는 하위 단어(sub-character word)를 의미한다.

[2] 본 논문에서 활용한 데이터는 https://github.com/naver/nlp-challenge 에서 확인할 수 있다.

시 각 모델 별 훈련 환경을 살펴볼 수 있다.

표 1. 선행 논문과 본 연구의 언어 모델 차이

	[8]	[9]	본 논문
언어	영어	한국어	한국어
BERT 모델	구글 버트	ETRI KorBERT	SNU KR-BERT
말뭉치	CoNLL	Korean prop-bank	NLP challenge

표 2. 한국어 사전 훈련 BERT 모델 [4]

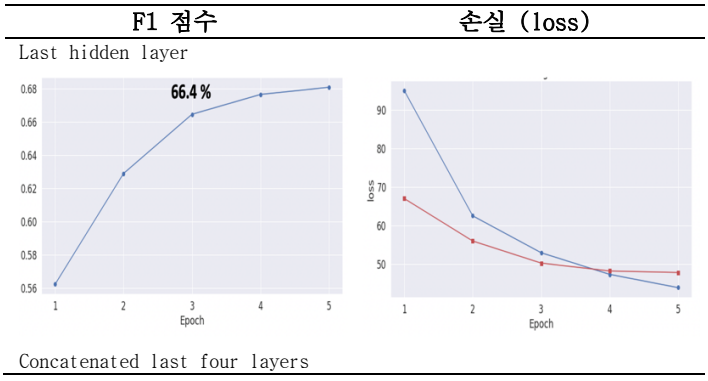
크기	ETRI KorBERT	SNU KR-BERT sub-character
단어	0.03 M	0.01 M
파라미터	109 M	0.96 M
훈련 데이터	23 GB	2.47 GB

NLP challenge 의미역 태깅 말뭉치는 약 33만 문장으로 구성되어 있으며, 약 27만 문장을 학습에 사용하였고, 약 7만 문장을 평가에 사용하였다. 한국어 서술어 자질 추출을 위해서 꼬꼬마 형태소 분석기를 이용하였다.

표 3과 그림 3은 NLP challenge 말뭉치를 이용한 한국어 의미역 결정 실험 결과이다. 일반적으로 언어 모델에서 사용하는 BERT 모델의 마지막 히든 레이어를 사용하였을 때, 약 66.4%의 성능을 보였다. SNU KR-BERT는 12개의 인코더 블록으로 구성되어 있다. 이 중에서 본 연구에서는 3가지 방식으로 BERT 모델의 히든 레이어를 활용하였으며, 결과는 다음과 같다. 첫 번째로, 마지막 4개의 레이어를 이었을 때(concatenated), F1 점수는 67.9% 였다. 두 번째로, 마지막 4개의 레이어를 합(sum)했을 때는 65.3%의 성능을 보였다. 마지막으로, 11번째 레이어를 선택했을 때 F1 점수는 68.1% 였다.

표 3. 선택된 BERT 레이어에 따른 언어 모델 성능

BERT 레이어	F1 점수
Last hidden layer	66.4%
Concatenated last four layers	67.9%
Sum last four layers	65.3%
Second-to-last layer	68.1%



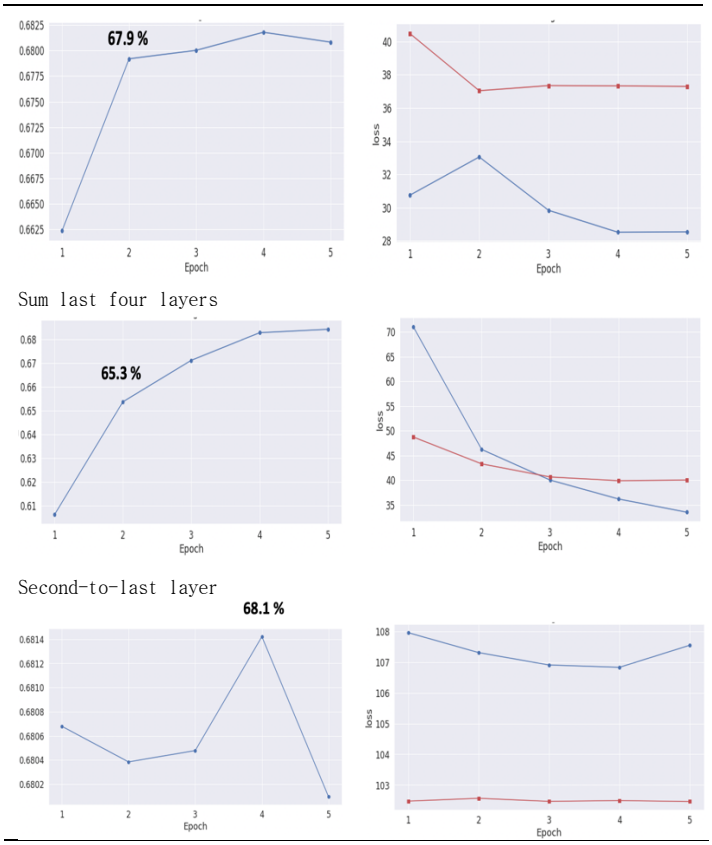


그림 3. 에폭 당 히든 레이어 별 F1 점수 및 손실

구현한 모델을 가지고 *'깨끗함을 최우선으로 내세우고 있기 때문이다.'* 라는 테스트 문장의 각 토큰 당 의미역 관계를 히트맵(heat map)으로 살펴보았다. 히트맵은 최종 레이블 값이 결정될 때 어떤 의미역이 중요한 역할을 했는지 시각화할 수 있는 기법이다. 가장 F1 점수가 낮은 마지막 4개의 히든 레이어를 합(Sum last four layers)한 언어 모델을 제외하고, 마지막 레이어(Last hidden layer)와 마지막 4개의 히든 레이어를 이은 것(Concatenated last four layers), 11번째 히든 레이어(Second-to-last layer)를 사용한 언어 모델을 활용해보았다. 표 4는 테스트 문장의 토큰 별 진리 의미역(ground-truth value)이다.

표 4. 토큰 별 의미역

토큰	진리 의미역	의미
깨끗함을	ARG1	목적어
최우선으로	ARGM-MNR	방법
내세우고	O	-
있기	O	-
때문이다	O	-

제안한 모델을 이용해서 표 4의 테스트 샘플의 레이블을 예측한 결과를 히트맵에 적용하여 나온 결과는 그림 4-6과 같다. 그림 4는 마지막 히든 레이어 언어 모델을 사용한 히트맵이다. 히트맵의 y축에 있는 것은 BERT 모델에 들어가는 토큰이며, x축에 있는 것은 의미역 종류이다. 이때 하나의 행에서 색이 진하면 진할 수록,

해당 의미역이 최종 레이블 결정에 중요한 역할을 했다는 것을 의미한다. 예를 들면, 토큰 *'깨'*에서 가장 진한 색은 O (의미역에 해당하지 않는 것), 그 다음으로 진한 것은 ARG1(목적어)이라는 것을 알 수 있으며, 토큰 *'최'*에서 가장 진한 색은 O, 그 다음으로 진한 것은 ARGM_MNR 이라는 것을 알 수 있다.

그림 5는 마지막 4개의 히든 레이어를 이은 언어 모델을 사용한 히트맵이다. 토큰 *'깨'*에서 가장 진한 색은 ARG3(술어로 기술된 행동의 도착점)이며, 토큰 *'최'*에서 가장 진한 색은 X(의미역에 해당하지 않는 것), 그 다음에 진한 색은 ARGM_EXT(범위)와 ARGM_PRP(대상과 같은 의미이거나 대상의 상태를 나타내면서 술어를 수식하는 것)이다.

그림 6은 11번 째 히든 레이어 언어 모델을 사용한 히트맵이다. 토큰 *'깨'*에서 가장 진한 색은 ARGM-LOC(장소)이며, 토큰 *'최'*에서 가장 진한 색은 ARG2 (술어로 기술된 행동의 시작점)이다.

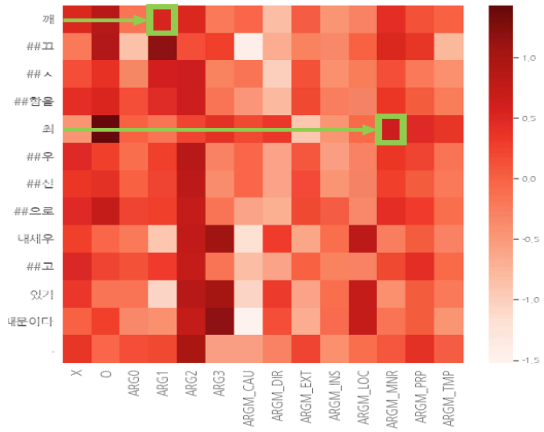


그림 4. Last hidden layer 히트맵

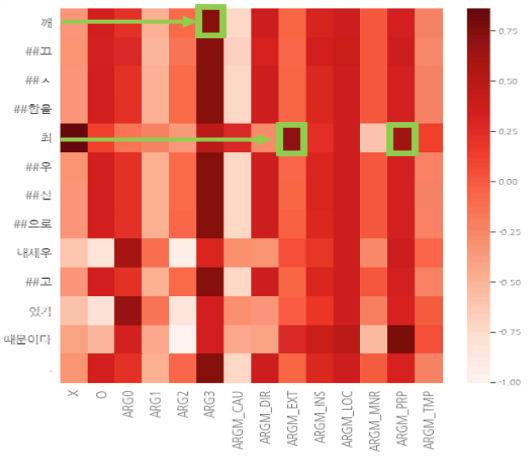


그림 5. Concatenated last four layers 히트맵

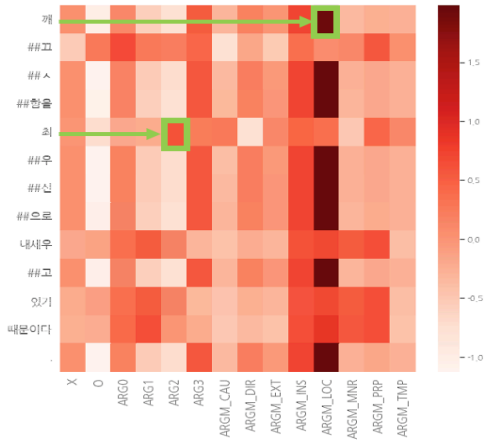


그림 6. Second-to-last layer 히트맵

6. 논의점

선행 연구에서는 Korean prop-bank를 주로 사용하였다[2,9]. Korean prop-bank는 문장 속 서술어와 그 서술어에 해당하는 의미역들이 정교하게 기술된 말뭉치이지만 의미역 결정을 위한 데이터 전처리 작업이 까다롭다. 반면에, 본 논문에서는 온라인에 공개된 NLP challenge 의미역 결정 말뭉치를 활용하였는데 해당 말뭉치에는 서술어에 대한 정보가 표기되어 있지 않다. 이러한 연유로 꼬꼬마 형태소 분석기를 통해 서술어에 대한 정보 자질을 부여했음에도 Korean prop-bank에 비해서 서술어 정보를 정확하게 표기하지 못하였기 때문에 모델의 성능을 높이는데 한계가 있었다.

본 논문에서는 BERT의 히든 레이어 중 어떤 레이어를 활용하는지에 따라서 달라지는 모델 성능에 좀 더 초점을 두었다. 흥미롭게도, 일반적으로 딥러닝 언어 모델에서 사용하는 마지막 히든 레이어를 사용했을 때보다 다른 히든 레이어를 선택했을 때 언어 모델의 성능이 더 높았다.

이러한 F1 점수를 토대로 언어 모델의 히트맵을 그려보면서, 각 토큰의 최종 레이블이 결정되기 전 어떤 의미역들이 큰 영향을 주었는지도 살펴보고자 하였다. 그 결과 마지막 4개의 히든 레이어를 이은 언어 모델(67.9%)과 11번 째 히든 레이어 언어 모델(68.1%)이 마지막 히든 레이어를 활용한 언어 모델(66.4%)보다 성능이 더 높았을 지라도 실제로 히트맵을 그려보니, last hidden layer를 활용한 언어 모델이 더 정확하게 의미역 판단을 한 것을 알아볼 수 있었다.

본 논문에서는 특정 과제에 따라서 어떤 히든 레이어를 취해야 하는지에 대한 연구가 활발히 이루어져야 한다고 주장한다. 이러한 과정을 통해서 딥러닝 모델을 언어학적으로 더 신뢰성있게 개발할 수 있으며, 더 나아가 딥러닝 모델의 언어 처리 방식 뿐만 아니라 사람의 언어 처리 방식에 대한 이해도까지도 높일 수도 있다고 기대한다.

감사의 글

이 논문은 2020년 대한민국 교육부와 한국연구재단의 일반공동연구지원사업의 지원을 받아 수행된 연구임(NRF-2020S1A5A2A03042760).

참고문헌

[1] 배장성, 이창기, 임수종. "Backward LSTM CRF 를 이용한 한국어 의미역 결정." *HCLT* (2015): 194-195.

[2] 이창기, 임수종, 김현기. "[우수논문] Structural SVM 기반의 한국어 의미역 결정." *한국정보과학회 학술발표논문집* (2014): 574 page

[3] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv*: 1810.04805 (2018).

[4] Lee, Sangah, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. "Kr-bert: A small-scale korean-specific language model." *arXiv preprint arXiv*:2008.03979 (2020).

[5] Marcheggiani, Diego, Anton Frolov, and Ivan Titov. "A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling." *arXiv preprint arXiv*:1701.02593 (2017).

[6] Roth, Michael, and Mirella Lapata. "Neural semantic role labeling with dependency path embeddings." *arXiv preprint arXiv*:1605.07515 (2016).

[7] Zhang, Yuhao, Peng Qi, and Christopher D. Manning. "Graph convolution over pruned dependency trees improves relation extraction." *arXiv preprint arXiv*:1809.10185 (2018).

[8] Peng, and Jimmy Lin. "Simple bert models for relation extraction and semantic role labeling." arXiv preprint arXiv:1904.05255 (2019).

[9] 배장성, 이창기, 임수종, 김현기. "BERT 를 이용한 한국어 의미역 결정." *정보과학회논문지* 47, no. 11 (2020): 1021-1026.

[10] Bae, Jangseong, Changki Lee, and Soojong Lim. "Korean Semantic Role Labeling using Backward LSTM CRF." *In Annual Conference on Human and Language Technology*, pp. 194-197. Human and Language Technology, 2015.

[11] Bae, Jangseong, and Changki Lee. "Korean Semantic Role Labeling using Stacked Bidirectional LSTM-CRFs." Journal of KIISE 44, no. 1 (2017): 36-43.

[12] 박광현, 나승훈, 신종훈, 김영길. "BERT 를 이용한 한국어 자연어처리: 개체명 인식, 감성분석, 의존 파싱, 의미역 결정." *한국정보과학회 학술발표논문집* (2019): 584-586.

[13] Li, Zuchao, Shexia He, Jiaxun Cai, Zhuosheng Zhang, Hai Zhao, Gongshen Liu, Linlin Li, and Luo Si. "A unified syntax-aware framework for semantic role labeling." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2401-2411. 2018.

[14] Došilović, Filip Karlo, Mario Brčić, and Nikica Hlupić. "Explainable artificial intelligence: A survey." In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 0210-0215. IEEE, 2018.

[15] Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. "A survey of methods for explaining black box models." *ACM computing surveys (CSUR)* 51, no. 5 (2018): 1-42.

[16] Lipton, Zachary C. "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue* 16, no. 3 (2018): 31-57.