KoBERT와 KR-BERT의 은닉층별 통사 및 의미 처리 성능 평가

최선주[0], 박명관, 김유희[+]
동국대학교, 신한대학교[+]
sunjoo@dongguk.edu, parkmk@dongguk.edu, euhkim@shinhan.ac.kr

How are they layerwisely 'surprised', KoBERT and KR-BERT?

Sunjoo Choi[0], Myung-Kwan Park, Euhee Kim[+]
Dongguk University, Shinhan University[+]

요 약

최근 많은 연구들이 BERT를 활용하여, 주어진 문맥에서 언어학/문법적으로 적절하지 않은 단어를 인지하고 찾아내는 성과를 보고하였다. 하지만 일반적으로 딥러닝 관점에서 NLL기법(Negative log-likelihood)은 주어진 문맥에서 언어 변칙에 대한 정확한 성격을 규명하기에는 어려움이 있다고 지적되고 있다. 이러한 한계를 해결하기 위하여, Li et al.(2021)은 트랜스포머 언어모델의 은닉층별 밀도 추정(density estimation)을 통한 가우시안 확률 분포를 활용하는 가우시안 혼합 모델(Gaussian Mixture Model)을 적용하였다. 그들은 트랜스포머 언어모델이 언어 변칙 예문들의 종류에 따라 상이한 메커니즘을 사용하여 처리한다는 점을 보고하였다. 이 선행 연구를 받아들여 본 연구에서는 한국어 기반 언어모델인 KoBERT나 KR-BERT도 과연 한국어의 상이한 유형의 언어 변칙 예문들을 다른 방식으로 처리할 수 있는지를 규명하고자 한다. 이를 위해, 본 연구에서는 한국어 형태통사적 그리고 의미적 변칙 예문들을 구성하였고, 이 예문들을 바탕으로 한국어 기반 모델들의 성능을 놀라움-갭(surprisal gap) 점수를 계산하여 평가하였다. 본 논문에서는 한국어 기반 모델들도 의미적 변칙 예문을 처리할 때보다 형태통사적 변칙 예문을 처리할 때 상대적으로 보다 더 높은 놀라움-갭 점수를 보여주고 있음을 발견하였다. 즉, 상이한 종류의 언어 변칙 예문들을 처리하기 위하여 다른 메커니즘을 활용하고 있음을 보였다.

주제어: KR-BERT, KoBERT, 언어 변칙, 놀라움-갭(surprisal gap)

1. 서론

최근 트랜스포머 모델에 관심이 높아지면서 이를 활용한 연구 역시 활발히 이루어지고 있다. 특히, 트랜스포머 모델은 자연어처리(NLP) 과제에서 괄목할만한 성과를 보여주었다. 대부분의 선행 연구들은 트랜스포머 언어모델의 언어 학습/처리를 살펴보기 위하여 다중선택 과제(multiple-choice task)을 활용하여 수행되어왔다. 즉, 언어모델은 일반적으로 주어진 문맥에서 적절하지 않다고 판단되는 단어보다 적절한 단어가 더 높은 우도(likelihood)를 부여하는 것이다. 하지만 우도는 단어들이 언어학/문법적인 측면에서 어떻게 서로 다른 성격을 갖는지를 구분하는 데는 어려움이 있다 [1]. 이러한 한계를 극복하기 위하여 최근 [2]는 언어모델의 언어처리 과정을 보다 더 구체적이고 상세하게 살펴보기 위하여 가우시안 혼합 모델(Gaussian Mixture Model)을 사용하였다.

언어 변칙(linguistic anomaly)은 여러 종류가 있다고 보고되었다. [3]은 형태통사론적 변칙과 의미적 변칙을 구분하였다(e.g., "colorless green ideas sleep furiously." vs. "furiously sleep ideas green colorless."). 또한, 심리언어학의 사건-관련전위(event-related potential, ERP) 기반 방법론에서는 언어 변칙의 종류에 따라 우리의 뇌가 다르게 반응한다고 본다. 다시 말해, 의미적 변칙 구문에선 N400 효과가 발견되며, 형태통사적 변칙 구문에선 P400 효과가 나타난다는 것이다. 즉, 상이한 종류의 언어 변칙에 따라 ERP 결과 패턴이 다르게 나타난다는 것을 말한다. 지금까지 수행되어온 이론적 혹은 실험언어학적 연구 결과를 받아들여 최근 [2]는 과연 언어모델도 언어 변칙의 종류에 따라 각 은닉층(hidden layer)에서 다른 '놀라움-갭(surprisal gap)' 현상을 보여주는지를 탐구하였다. 그들은 BERT [5], RoBERTa [6], 그리고 XLNet [7]의 은닉층에서의 놀라움(surprisal)을 측정하기 위하여 밀도 추정(density estimation)을 활용하였다. 결론적으로 언어모델도 내부적으로 언어 변칙의 종류를 구분한다는 점을 발견하였다.

[2]의 연구 결과를 받아들여 본 논문에서는 사전학습(pre-trained)된 한국어 기반 언어모델인 KR-BERT [8]와 KoBERT[1]를 파인튜닝(fine-tuning)하여 상이한 종류의 한국어 변칙 예문이 주어졌을 때, 이를 인지하고 다르게 처리하는지 살펴보고자 한다. 이를 위해, 선행 연구를 참고하여 언어 변칙 종류를 형태통사적 변칙과 의미적 변칙 두 가지로 나누었다. 그리고 가우시안 혼합 모델을 KR-BERT와 KoBERT의 은닉층에서의 각 층에 맞도록 훈련시켰다. 결과적으로, 한국어 기반 모델 역시 기존 영어 기반 트랜스포머 언어모델이 보여주었던 것과 마찬가지로 상이한 종류의 언어 변칙에 따라 처리 양상을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 실험 구성을 제시한다. 4장에서는 한국어 기반 모델이 어떻게 각 은닉층에서 언어 정보

1) https://github.com/SKTBrain/KoBERT.

를 인지/처리하는지에 관해 결과를 보고한다. 5장에서는 결론에 대해 기술한다.

2. 관련연구

BERT 모델이 나온 후, 많은 연구들이 언어모델에서 어떻게 언어 정보가 처리되는지에 관하여 언어모델의 은닉층에 문장의 표상을 입력시켜 탐구하였다. [9]는 언어 과제의 수행에서 각 은닉층의 기여도를 살펴보기 위하여 '끝머리 탐침(edge probing)' 기법을 이용하였고, 두 가지 주목할 결과를 도출하였다. 즉, 중간 은닉층은 통사적 과제를 수행할 때 상대적으로 더 높은 기여하였고, 더 상위의 은닉층일수록 의미적 과제의 수행에 기여를 한다는 것이다. [10]은 BERT의 중간 은닉층에서 직접 혹은 간접 목적어 구문들을 잘 구별한다는 점을 보고하였다. 많은 선행 연구들이 공통적으로 실험을 통하여 BERT의 중간 은닉층이 일반적으로 통사적 정보를 처리하는 구간이라는 결과를 보고하였다. 선행 연구에서 더 나아가 [2]는 언어 변칙 예문들을 바탕으로 각각의 다른 은닉층마다 특정한 언어 정보와 연관되어 있는지 탐구하고자 하였다. 이를 바탕으로, 언어모델 역시 상이한 종류의 언어 변칙에 따라 다른 메커니즘을 이용하여 처리할 수 있다는 점을 발견하였다. 다음 장에서 [2]의 연구를 자세하게 살펴보고자 한다.

2.1 Li et al. 2021

[2]의 연구에서는 사전학습 모델인 BERT, RoBERTa, 그리고 XLNet 세 가지 트랜스포머 모델을 파인튜닝하였다. 각 모델의 구성은 문맥을 반영한 12개의 은닉층과 입력을 위한 1개의 은닉층이 있으며. 모든 단어들은 768 차원의 임베딩 벡터가 되어 사전학습 모델의 입력으로 사용된다. The British National Corpus [11]에서 무작위로 문장들을 추출하여 가우시안 혼합 모델을 훈련시켰으며, BLiMP [12]로 평가하였다. 실험 데이터 구성으로서 언어 변칙의 종류를 다음과 같이 두 가지 기준으로 나눴다.

1. 형태통사적 변칙: 단어의 굴절 형태의 오류, 혹은 잘못된 동사 시제 (e.g., The cats won't eat the food that Mary gives them. vs. *The cat won't eating the food that Mary gives them)
2. 의미적 변칙: 유생성(animacy)과 같은 선택제약(selectional restriction) 위반. (e.g., The cat won't eat the food that Mary gives them vs. #The cat won't bake the food that Mary gives them)

각 문장의 대립쌍은 정문과 언어 변칙 문장으로 구성된다. 각 은닉층에서는 표준편차에 따라 정문과 언어 변칙 문장의 놀라움 점수의 차이를 놀라움-갭으로 정의하였다.

[2]의 실험결과 가장 성능이 좋은 모델은 RoBERTa이며, RoBERTa모델은 의미적 변칙 보다는 형태통사론적 변칙 구문을 처리할 때 하위 은닉층에서 통계적으로 유의미한 놀라움이 나타났다고 보고하였다. 이 결과는 언어모델이 언어 변칙 종류의 차이를 인지하여 다르게 처리한다는 것을 의미한다. 요약 결과는 다음 그림 1과 같다.

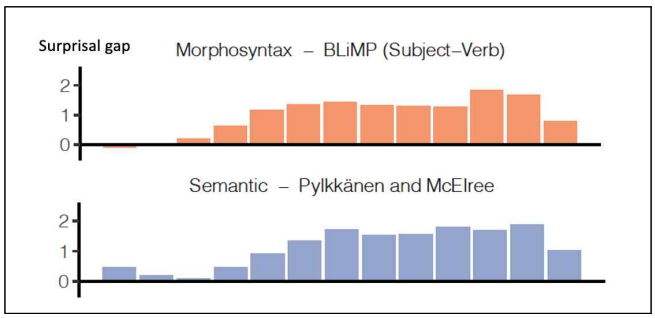


그림 1. 영어 기반 RoBERTa 모델에서 나타난 두 가지 과제의 은닉층별 놀라움 갭 패턴

3. 실험

3.1 한국어 데이터

[2]의 데이터 구성을 참고하여 본 연구에서도 두 가지 종류로 한국어 변칙을 나누었다. 형태통사적 변칙은 [2]의 선례를 참고하여 견본을 만들어 구성하였고, 의미적 변칙은 기존 심리언어학 실험 연구의 데이터 구성을 기반으로 만들었다 [13][14][15][16]. 다음 표 1에서 본 연구의 대표 실험 예문의 대립쌍을 정리하였다.

표 1. 실험 예문 구성

언어 변칙 종류	문장 대립쌍
형태통사	(1) a. 아이가 밥을 **먹었다**. b. *아이가 밥을 **먹으시었다**. (2) a. 엄마가 음식을 **요리하고** 있다. b. *엄마가 음식을 **요리하지** 있다. (3) a. 선생님이 그 학생들을 **세 명을** 만났다. b. *선생님이 그 학생을 **세 명을** 만났다.
의미	(1) a. 강아지가 공을 굴리며 **짖었다**. b. #강아지가 공을 굴리며 박수를 **쳤다**. (2) a. 엄마가 음식을 **했다**. b. #엄마가 음식을 **때렸다**.

표 1에서 정리한 언어 변칙 유형에 따라 270개의 실험 문장을 구성하였고, 각 문장 세트는 정문과 변칙 문장의 대립쌍으로 이루어졌다.

3.2 놀라움-갭(Surprisal gap)

본 연구에서의 놀라움 점수는 목표 문장의 모든 토큰

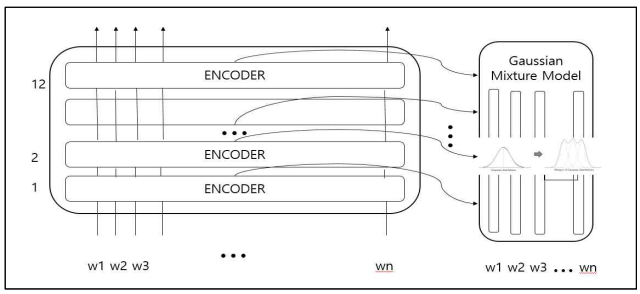
의 놀라움에 대한 평균값으로 계산하였다. 그리고 정문과 언어 변칙 문장의 차이를 놀라움-갭 점수로 다음과 같은 공식을 바탕으로 계산하였다.

$$\text{surprisal gap}_L(\mathcal{D}) = \frac{\mathbb{E}\{\text{surprisal}_L(\boldsymbol{s}_i') - \text{surprisal}_L(\boldsymbol{s}_i)\}_{i=1}^{n}}{\sigma\{\text{surprisal}_L(\boldsymbol{s}_i') - \text{surprisal}_L(\boldsymbol{s}_i)\}_{i=1}^{n}}$$

3.3 한국어 BERT을 이용한 변칙 문장 분석 모델

본 논문에서는 사전학습된 한국어 기반 BERT를 변칙 문장 분석에 적용한다. 본 논문의 분석 모델은 그림 2와 같이 한국어 기반 BERT의 기본 구조인 트랜스포머의 인코더 12개 은닉층으로부터 각 토큰별 문맥 임베딩 벡터를 추출한다. 다음으로, 각 입력 문장에 대해 추출한 토큰 임베딩 벡터를 사용하여 가우시안 혼합 모델로 훈련한다.

그림 2와 같이 분석 모델의 구조는 Hugging Face Model Library[2]에서 제공하는 사전훈련된 KoBERT 그리고 KR-BERT의 베이스 버전을 사용한다. 형태통사적 또는 의미적 변칙 문장을 언어모델의 은닉층별로 식별하기 위해, 한국어 기반 BERT의 은익층별로 임베딩 벡터를 추출하여 가우시안 모델링에 적합한 가우시안 혼합 모델(GMM)을 사용하였다. 모델을 훈련시키기 위해 국립국어원 말뭉치(Korean Corpus)에서 무작위로 선택한 문장들을 사용하였다. 그리고 3.1장에서 언급한 자료들을 바탕으로 구성된 실험 문장의 대립쌍(Test Corpus)을 모델 평가에 사용하였다. 학습된 GMM 언어 변칙 모델(GMM Anomaly Model)을 사용하여 형태통사적 그리고 의미적 변칙 과제별로 주어진 문장의 대립쌍에 대해 놀라움-갭과 모델의 정확도를 평가하였다. 그리고 만약 변칙 문장의 놀라움 점수가 정문의 것보다 높다면 문장의 대립쌍은 적절히 식별된다고 간주한다.



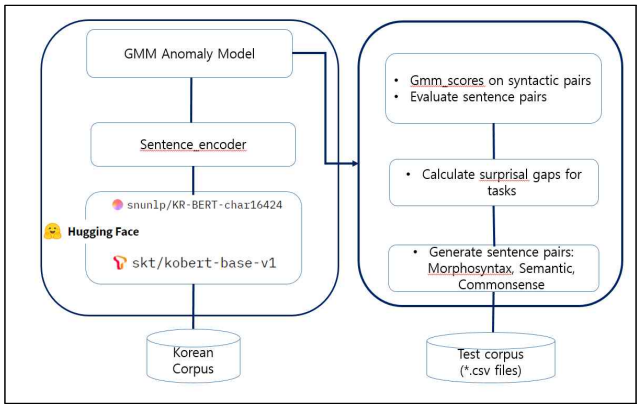


그림 2. 한국어 기반 BERT을 이용한 GMM 언어 변칙 모델

본 연구에서 GMM 언어 변칙 모델은 입력 은닉층(0)과 한국어 기반 BERT 언어모델의 12개의 은닉층(1~12)을 구성된다. 그림 3과 같이 모델의 변칙 과제 관점에서 보면 KR-BERT가 KoBERT보다 변칙 문장 분석에 대한 정확도가 높은 것으로 나타났다. 모델의 은닉층별 관점에서 보면 KR-BERT는 10번 은닉층이 가장 높은 정확도를 보였으며, KoBERT는 4번 은닉층의 정확도가 가장 높은 것으로 확인되었다.

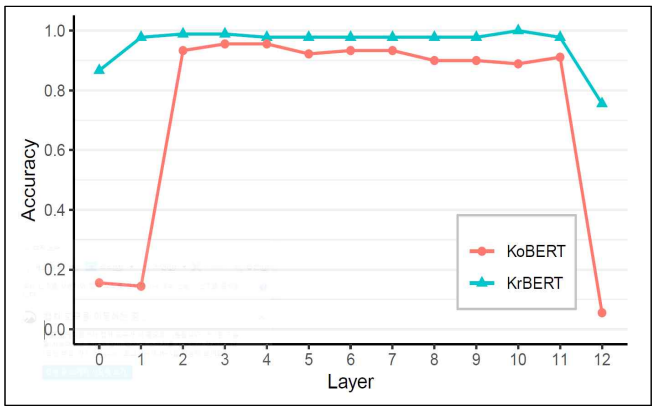


그림 3. 은닉층별 한국어 기반 BERT 정확도

추가적으로 본 연구에서는 각 토큰에서의 놀라움 점수와 토큰 출현 빈도 사이에서의 피어슨 상관 계수(Pearson correlation)를 계산하였다.

2) https://huggingface.co/models

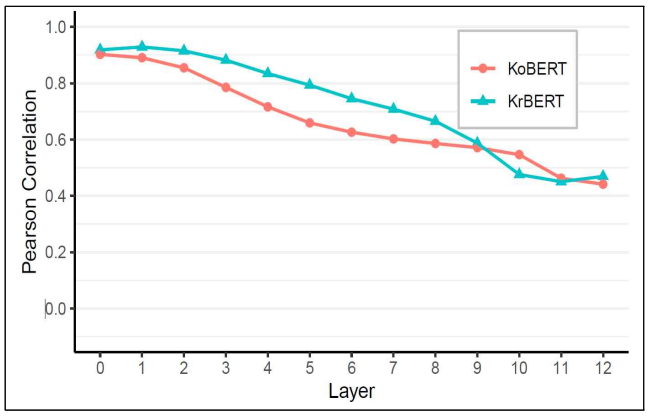


그림 4. 토큰 레벨 놀라움 점수와 출현 빈도수 사이의 피어슨 상관 계수

그림 4와 같이 KR-BERT가 KoBERT보다 변칙 문장 분석에 대한 토큰 레벨 놀라움 점수와 출현 빈도수 간의 상관관계가 거의 모든 은닉층에서 높다는 것을 알 수 있다. 두 모델에서 놀라움 점수와 출현 빈도 사이에 높은 상관관계가 나타났지만, 은닉층이 높아질수록 이 값은 점차적으로 줄어들었다. 하지만, 양의 상관관계는 마지막 은닉층까지 지속되었다는 점을 확인할 수 있다.

4. 실험 결과

본 논문에서는 선행 연구 [2]에서처럼 변칙 예문들을 바탕으로 각각 한국어 기반 BERT 언어모델의 은닉층마다 어떤 언어 정보가 포함되어 있는지 탐구하였다.

그림 5을 보면 KR-BERT와 KoBERT 모델을 구현하였을 때 두 개의 대표 예문에서 나타난 밀도 추정 값을 나타낸 것이다. 그림에서 어두운 색상일수록 더 높은 놀라움 값을 의미한다. 두 언어모델에서 (위 그림: KR-BERT, 아래 그림: KoBERT) 형태통사적 변칙 문장은 하위의 은닉층에서 높은 놀라움 점수를 보여주었고, 의미적 변칙 문장은 상대적으로 상위의 은닉층에서 반응이 나타나기 시작하였다. 이 실험 결과의 패턴은 선행 연구에서 보고하였던 결과와 일치하다 [9].

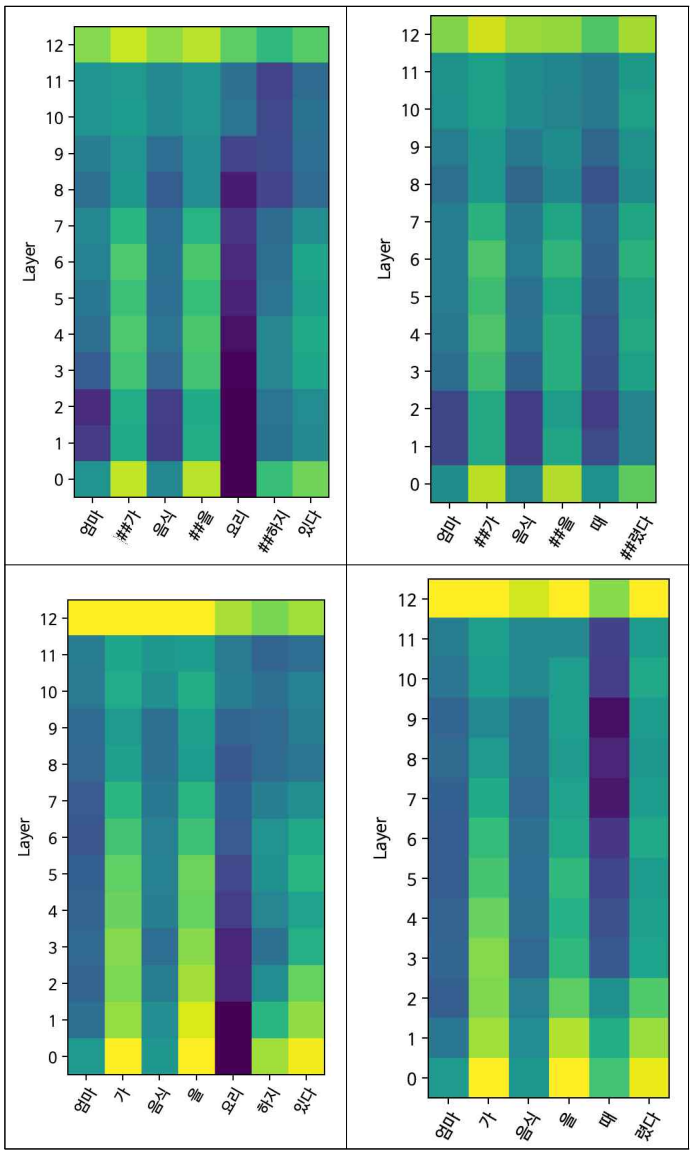


그림 5. KR-BERT(위)와 KoBERT(아래)를 이용한 형태통사적 변칙 예문(왼쪽)과 의미적 변칙 예문(오른쪽)의 밀도 추정

그림 6과 같이 언어 변칙의 종류에 따라 언어모델의 은닉층별로 놀라움-갭 분포도 다르게 나타난 것을 발견하였다. 추가적으로 언급할 흥미로운 결과는 두 언어모델이 보여주는 놀라움-갭 값의 분포의 차이이다. KR-BERT 모델에선 형태통사적 변칙 예문들이 상대적으로 높은 놀라움-갭 값을 2번부터 11번 은닉층까지 보여주었지만, 의미적 변칙은 상대적으로 낮은 놀라움-갭 값을 모든 은닉층에서 일정한 패턴으로 지속적으로 보여주었다. 하지만 KoBERT 모델에선 형태통사적 변칙 예문들은 3번에서 11번 은닉층까지 상대적으로 높은 놀라움-갭 값이 나타났으며, 의미적 변칙 예문들은 특히 4번에서 7번 은닉층에서 상대적으로 낮은 놀라움-갭 값이 나타났다.

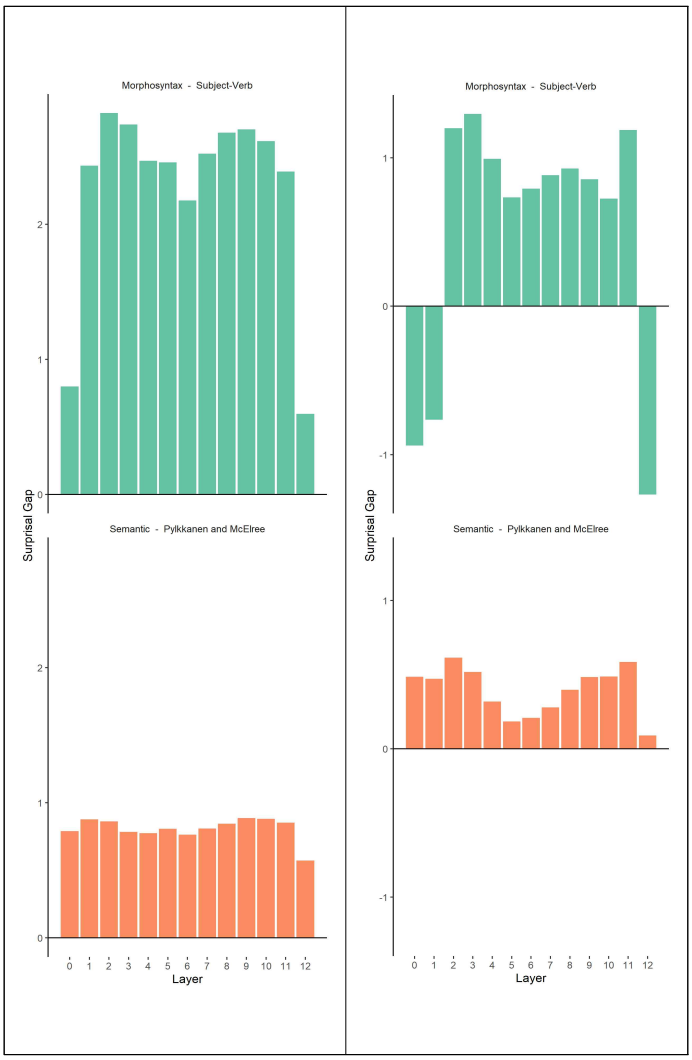


그림 6. 은닉층별 놀라움-갭 (KR-BERT 왼쪽, KoBERT 오른쪽)

본 연구에서는 한국어 기반 BERT의 은닉층별로 출현빈도가 높은 토큰들의 군집 분포의 양상을 살펴보았다. 다음 그림 7은 주성분 분석(principal component analysis, PCA)을 이용하여 국립국어원 코퍼스에서 무작위로 추출한 문장들을 두 개의 언어모델 KR-BERT과 KoBERT에 넣고 각 은닉층마다 토큰 빈도수의 분포를 차원 축소하여 나타낸 결과이다.

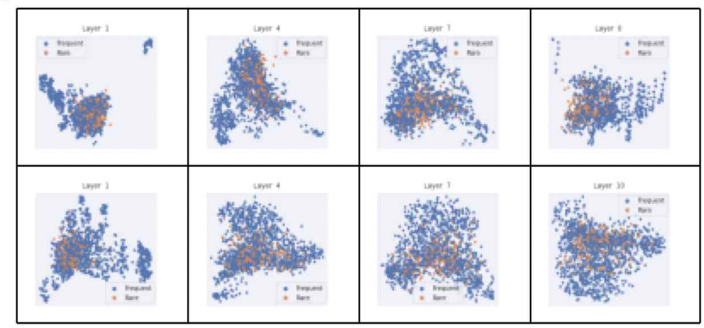


그림 7. KR-BERT (위)와 KoBERT (아래)에서 1, 4, 7, 10번 은닉층에서의 PCA 분포

그림 7에서 볼 수 있듯이 “드문”의 의미는 최소 20%의 빈도수 토큰을 말하며, “빈번”의 의미는 나머지 80%를 말한다. 두 가지 케이스에서 모두 1번, 4번, 7번, 10번 은닉층에서 드문 토큰과 빈번한 토큰이 다른 구역을 차지한 점을 확인할 수 있었다. 여기서 KR-BERT가 KoBERT보다 주어진 4개의 은닉층에서 구역이 좀 더 밀집되어 있는 것을 볼 수 있다.

5. 결론

한국어에서 나타난 언어 변칙의 종류에 따라 한국어 기반 언어모델인 KR-BERT와 KoBERT는 은닉층에서 다르게 처리하는 패턴을 보여주었다. 본 연구에서 관찰된 양상들은 중간 은닉층에서 주로 통사 정보를 인지/처리하고, 상위 은닉층에서 주로 의미 정보를 인지/처리한다는 선행 연구 Li et al.(2021)의 결과와 다르다. 즉, 본 연구에는 형태통사적 변칙 예문들은 의미적 변칙 예문들보다 전체 은닉층에 걸쳐서 상대적으로 더 높은 놀라움-갭 값을 보여주었다. 한국어 기반 BERT와 영어 기반 BERT의 이와 같은 언어 처리적 차이는 토큰나이저 방식에 기인한다. 즉, 문장을 띄어쓰기 단위로 분절하면 단어 토큰이 구분되는 영어와 달리, 한국어는 교착어이기 때문에 띄어쓰기만으로는 단어 토큰을 구분하기 어려워서 서브워드 분절 방식을 사용한다. 다시 말해, 토큰 분절 방식이 언어 처리에도 영향을 주고 있다고 추정된다. 결론적으로, 본 연구의 실험 결과를 통하여 한국어 기반 트랜스포머 언어모델이 각 은닉층에서의 언어 정보를 인지하기 위하여 상이한 방법으로 두 가지 변칙 예문들을 처리한다는 점을 보여주었다.

감사의 글

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

참고문헌

[1] Gulordava, Kristina, et al. Colorless green recurrent networks dream hierarchically. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018.

[2] Li, Bai, et al. How is BERT surprised? Layerwise detection of linguistic anomalies, 2021.

[3] Chomsky, Noam. Syntactic Structures, Mouton and Co. 1957.

[4] Kutas, Marta, et al. Psycholinguistics electrified II, Academic Press, 2006.

[5] Jacob, Devlin, et al. BERT: Pre-training of deep bidirectional transformers for language understanding, 2019.

[6] Liu, Yinhan, et al. RoBERTa: A robustly optimized BERT pretraining approach, 2019.
[7] Yang, Zhilin, et al. XLNet: Generalized autoregressive pretraining for language understanding, 2019.
[8] Lee, Sangah, et al. KR-BERT: A small-scale Korean-specific language model. 2020.
[9] Tenney, Ian, et al. BERT rediscovers the classical NLP pipeline, 2019.
[10] Kelly, M. Alex, et al. Which Sentence Embeddings and Which Layers Encode Syntactic Structure? 2020.
[11] Leech, Geoffrey Neil. 100 Million words of English: the British National Corpus (BNC). Language Research, 28:1-13. 1992.
[12] Warstadt, Alex, et al. BLiMP: The benchmark of linguistic minimal pairs for English. Transactions of the Association for Computational Linguistics, 8:377-392. 2020.
[13] Warren, Tessa, et al. Comprehending the impossible: what role do selectional restriction violations play? Language, cognition and neuroscience, 30(8): 932-939. 2015.
[14] Lee, Osterhout and Mobley, Linda. Event-related brain potentials elicited by failure to agree. Journal of Memory and language, 34(6):739-773. 1995.
[15] Lee, Osterhout and Nicol, Janet. On the distinctiveness, independence, and time course of the brain responses to syntactic and semantic anomalies. Language and cognitive processes, 14(3):283-317. 1999.
[16] Pylkkänen, Liina and McElree, Brian. An MEG study of silent meaning. Journal of cognitive neuro-science, 19(11):1905-1921. 2007.