

언어학 관점에서의 한국어 대조학습 기반 문장 임베딩의 허위 문맥화에 대한 고찰

정유현⁰, 한명수, 채동규

한양대학교 인공지능학과
{myngsoo, robo0725, dongkyu} @ hanyang.ac.kr

Analyzing Spurious Contextualization of Korean Contrastive Sentence Representation from the Perspective of Linguistics

Yoo Hyun Jeong, Myeongsoo Han, Dong-Kyu Chae
Dept. of Artificial Intelligence, Hanyang University

요약

본 논문은 사전 학습 언어 모델의 특성인 이방성과 문맥화에 주목하여 이에 대한 분석 실험과 한국어 언어 모델만의 새로운 관점을 제안한다. 최근 진행된 영어 언어 모델 분석 연구에서 영감을 받아, 한국어 언어 모델에서도 대조학습을 통한 이방성과 문맥화의 변화를 보고하였으며, 다양한 모델에 대하여 토큰들을 문맥화 정도에 따라 분류하였다. 또한, 한국어의 언어학적 특성을 고려하여, 허위 문맥화를 완화할 수 있는 토큰을 문맥 중심어로, 문맥 중심어의 임베딩을 모방하는 토큰을 문맥 기능어로 분류하는 기준을 제안하였다. 간단한 적대적 데이터 증강 실험을 통하여 제안하는 분류 기준의 가능성을 확인하였으며, 본 논문이 향후 평가 벤치마크 및 데이터셋 제작, 나아가 한국어를 위한 강건한 학습 방법론에 기여하길 바란다.

주제어: 대조 학습, 언어학, 허위 문맥화

1. 서론

문장 임베딩 (sentence embedding)을 개선할 수 있는 대표적인 방법론에는 대조학습 (contrastive learning)이 있다. 특히나 자연어 처리 분야에서는 SimCSE [15]가 대조학습을 통해 성능 개선과 기준 사전학습 모델의 다양한 문제점을 해결하였다. SimCSE는 드롭아웃 (dropout)을 기반으로 한 비지도 대조쌍을 생성하여 자연어 처리만의 새롭고 강건한 (robust) 대조학습을 제안하였다.

한편, 영어 기반 사전학습 언어모델의 문맥화 (contextualization) 및 표현 공간(representation space)의 이방성 (anisotropy)에 관한 분석 연구들이 최근 수 행된 바 있다 [12, 17]. 문맥화란 각 토큰 (token)들이 다른 문장(문맥)에서 다른 의미를 갖고 있음을 의미하며, 이방성은 표현 공간의 임베딩 벡터들이 균일하게 분포되어 있지 않음을 뜻한다. 참고문헌 [12]에서는 여러 지표들을 도입하여 언어 모델들을 분석하였으며, 실험적으로 사전 학습 언어 모델의 문맥화 특성과, 표현 공간의 이방성을 보였다. SimCSE는 대조학습이 표현 공간을 균등하게 만들어 이방성의 문제를 해결할 수 있음을 주장하였다. 그러나, 이방성 문제의 해결이 실제 성능 개선으로 이어질 수 있음을 확인하지 못하였으며, 나아가 대조학습과 실제 문맥화 지표와의 연관성에 대한 의미 있는 해석은 부족하였다.

참고문헌 [17]에서는 영어 기반 사전학습 언어모델의 대조학습 관련 문맥화 문제에 대한 분석을 진행하였다. 해당 논문에서 대조학습이 빈도수 편향 [16]을 해결하고,

문맥화와 관련이 있음을 보였다. 특히 토큰들의 허위 문맥화 (spurious contextualization)의 개념을 제안하였는데, 이는 토큰들이 실제로는 뜻이 변하지 않음에도 불구하고 문맥화된 것을 의미한다. 나아가 허위 문맥화가 일어나는 패턴을 발견하였으며, 허위 문맥화를 일으키는 단어를 기능어로, 이를 완화하는 토큰들을 중심어 또는 의미어로 분류하였다.

이러한 연구 결과들에 영감을 받아, 본 논문은 한국어 언어 모델에서의 대조학습에 따른 문장 임베딩의 변화를 분석하고자 한다. 특히 영어와 한국어의 언어적 특성이 다르다는 것에 집중하여, 우리가 발견한 분석 결과들과 기준의 분석 결과의 공통점과 차이점을 분석한다. 또한 다양한 분석 결과들을 바탕으로 한국어 자연어 처리 모델에서의 허위 문맥화 관련 토큰들을 언어학을 기반으로 분류한다. 이렇게 만들어진 분류 기준을 바탕으로 대조학습이 진행된 한국어 언어 모델에 적대적 (adversarial) 데이터 증강을 진행하여 결과를 확인한다.

본 논문의 기여는 다음과 같다.

- 대조학습을 한 한국어 사전 학습 언어 모델의 표현 공간의 특성을 분석하였고, 특징이 있는 토큰들을 찾았다.
- 허위 문맥화와 관련된 토큰들을 한국어의 언어적 특성을 고려하여 정리하였으며, 이를 문맥 중심어와 문맥 기능어로 분류하였다.
- 분류된 토큰들을 기반으로 적대적 데이터 증강 방법론을 제안하였으며, 향후 벤치마크 연구 및 모델의 강건성 연구에 활용될 가능성을 보았다.

2. 관련 연구

언어 모델: Transformer 모델 [8]의 등장 이후, 언어 모델은 많은 발전을 이뤄냈다. 대표적으로 BERT [10]와 RoBERTa [11]는 좋은 성능을 내는 기초 모델로 사용된다. 한국어 언어 모델의 경우에도 이를 바탕으로 만들어졌으며, KLUE [14] 등의 한국어 데이터셋에서 학습한 모델들이 존재한다.

이방성과 문맥화: 사전 학습된 언어 모델의 토큰 임베딩의 이방성 문제는 실험적으로 증명되었다. 즉 이러한 모델의 표현 공간은 균방성(isotropy)을 가지지 않으며, 좁은 원뿔 형태를 가지고 있음이 밝혀졌다 [12]. 이방성의 문제가 미세 조정 학습을 통해 해결될 수 있음을 보인 논문은 존재하지만 [16], 이것이 이방성 문제의 해결을 통해 이뤄진 결과임을 입증하진 못했다.

문장 임베딩을 대조학습을 통해 개선하고자 한 연구는 SimCSE [15]가 대표적이다. 해당 논문에서는 대조학습을 통한 성능 향상과 더불어, 주장하는 방법론이 균등한 표현 공간을 형성하는데 이바지함을 정량적 지표들을 통해 보였다. 최근, 대조학습이 이방성 문제 해결에 직접적인 역할을 할 수 있음을 분석한 연구가 진행되었다 [17]. 해당 연구에서는 다양한 문맥화 관련 지표를 활용하여 대조학습의 표현 공간의 기하학적 특성과, 특정 의미 있는 토큰들이 문장 임베딩을 결정하는 허위 문맥화의 개념을 주장하였다.

3. 예비 연구

본 장에서는 참고문헌 [17]에서 영감을 받아, 분석 실험을 진행하기 전 한국어 토큰에 대한 예비 연구를 수행하여 결과를 보고한다. 참고문헌 [12]에서 제안된 여러 지표들이 사용되었으며, 각각은 아래와 같다.

3.1 분석 실험을 위한 지표

Self-Similarity: 말뭉치 (corpus) 내 서로 다른 문장들에서의 토큰 표현의 유사도를 측정한다. 다른 문장은 문맥이 다르다고 가정하며, 때문에 높은 self-similarity는 문맥화가 되지 않았음을 의미한다. 토큰을 x , 코사인 유사도 (cosine similarity)를 \cos 라 했을 때, 아래의 식 1처럼 모든 문장에 대한 토큰 x 의 코사인 유사도의 평균을 통해 self-similarity를 계산할 수 있다.

$$\text{selfsim}(x) \triangleq \text{empirical_mean}(\cos(u, v)). \quad (1)$$

Intra-sentence Similarity: 같은 문맥 내의 토큰 유사도를 측정하는 지표이다. 아래의 식 2처럼, 문장 s 내의 모든 토큰들의 문장 임베딩을 mean-pooling하고, 각 토큰의 임베딩 x 와 mean-pooling 임베딩의 평균을 구한다. Intra-sentence similarity는 토큰 표현이 의미 공간 (semantic space)에서 얼마나 수렴하는지를 판단하는 지표로 사용될 수 있다.

$$\vec{s} \triangleq \frac{1}{n} \sum \vec{x}, \\ \text{intrasim}(\vec{x}) = \frac{1}{n} \sum \cos(\vec{s}, \vec{x}). \quad (2)$$

Anisotropy Baseline: 참고문헌 [17]과 같이, 앞선 두 유사도 지표를 표현 공간의 이방성을 고려하여 일반적인 분포에 알맞게 조절하기 위한 anisotropy baseline을 도입한다. Anisotropy baseline은 무작위로 뽑은 다른 문맥에서의 토큰들의 코사인 유사도의 평균으로 구한다. 앞선 두 지표에 anisotropy baseline 값을 빼주어 최종적으로 분석 실험에 활용한다.

3.2 빈도수 편향과 Self-Similarity

영어 기반 BERT 모델에서는 말뭉치 내 등장하는 단어 빈도수의 불균형에 따른 임베딩의 이방성이 확인되었다 [16]. 이러한 임베딩의 이방성을 언어 모델이 문맥화가 되었다고 보는 입장이 존재하였으며 [12], 이것이 이전 통계 기반 언어 모델과 Transformer 기반의 모델의 차이로 간주되었다. SimCSE [15]는 사전 학습된 모델 (BERT, RoBERTa)에 추가적인 대조학습을 진행하여 이방성의 문제를 해결하고자 하였다.

그러나, SimCSE의 대조학습으로도 빈도수로 인한 이방성 문제는 완전히 해결되기 어려웠다. [17]은 학습 전후의 self-similarity의 차이를 측정하여 대조 학습으로 빈도수의 문제를 해결될 수 있는지 검증하였다. self-similarity의 차이는 이방성 baseline을 고려하여 조정되었다. 이 때 self-similarity를 ss , 이방성 baseline을 ani , 대조 학습 전후로 각 지표를 명시하였으며, self-similarity의 차이는 참고문헌 [17]과 같이 SSC라 명명한다. SSC는 아래의 식 3처럼 계산된다.

$$ssc = (ss_{\text{학습후}} - ani_{\text{학습후}}) - (ss_{\text{학습전}} - ani_{\text{학습전}}). \quad (3)$$

본 논문은 참고문헌 [17]에서 발견된 결과를 한국어 사전 학습 언어 모델에서도 찾을 수 있는지 확인하기 위해 다음과 같은 모델들, 데이터셋, 그리고 대조학습을 사용하였다. 모델은 영어 모델에서 강건함을 보인 BERT와 RoBERTa를 KLUE [14] 기반으로 학습한 KLUE-BERT와 KLUE-RoBERTa를 huggingface를 통해 다운로드 하였다.¹ 대조학습 훈련 방법은 SimCSE [12]에서 착안하여 비지도, 지도 (supervised), 멀티 태스크 (multi-task)를 사용하였다. 각 학습 방법을 위한 데이터셋은 Korean Wikipedia corpus와 KorNLI [12]을 사용하였다. 대부분의 실험 환경은 SimCSE 및 기존에 검증된 방법론²을 따랐다.

빈도수 편향을 측정하기 위한 데이터셋은 대조학습 때 사용되지 않았던 KorSTS의 테스트 데이터셋과 KorNLI의 검증 데이터셋이다. KorSTS의 테스트 셋은 sentence1만을 사용하여 문맥이 중복되는 것을 피하였고, KorNLI의 경우 premise만을 사용하였다. 각 데이터셋에 대하여 빈도수

¹ <https://huggingface.co/klue/bert-base>, <https://huggingface.co/klue/roberta-base>

² <https://github.com/BM-K/Sentence-Embedding-is-all-you-need>

를 분석 실험 전에 측정하였으며, 빈도수가 낮은 단어의 경우 적절한 문맥 지표를 측정하기 어렵기 때문에 KorSTS의 경우 5번 이상, KorNLI의 경우 10번 이상 등장한 토큰들만을 사용하였다. 전반적인 분석 실험용 데이터셋에 대한 상세 내용은 표 1과 같다.

데이터셋	종류	문장수	토큰수
KorSTS	Test	1380	580
KorNLI	Validation	1570	476

표 1. 분석 실험에 사용된 데이터셋에 대한 상세 설명.

비지도 대조학습		지도 대조학습		멀티 태스크	
SS (↓)	SS (↑)	SS (↓)	SS (↑)	SS (↓)	SS (↑)
0 그러나	-	##고	흑백	##고	16
1 [CLS]	12	명	목요일	##어요	13
2 ##한다	##프	##한다	폭발	##한다	방
3 ##됨	##2	##을	기관	명	목요일
4 ##어요	11	##습	방	##를	25
5 ##이다	##妣	##어요	작년	##을	수요일
6 [SEP]	[UNK]	##이	투표	##가	폭발
7 ##합니다	##비	##하	승리	##하고	9
8 주당	##y	##를	금요일	##하	승리
9 .	소	##가	신문	##됨	금요일

표 2. KorSTS 데이터셋에 대하여, KLUE-BERT를 SimCSE로 학습한 모델들 (비지도, 지도, 멀티 태스크)의 훈련 전후 토큰들의 Self-Similarity의 변화.

비지도 대조학습		지도 대조학습		멀티 태스크	
SS (↓)	SS (↑)	SS (↓)	SS (↑)	SS (↓)	SS (↑)
0 그러나	소	##습	석방	##합니다	석방
1 [SEP]	석방	##합니다	소	.	소
2 [CLS]	16	##고	오바마	##습	투표
3 .	12	##가	투표	##가	오바마
4 ##다고	11	##를	흑백	##고	9
5 ##하기	투표	.	우크라이나	##한다	11
6 ##으며	충돌	##한다	무리	##다	우크라이나
7 현재	##비	##을	국제	##니다	12
8 또한	네	##하	붕괴	##를	16
9 ##습	##터	##이	금요일	##하	10

표 3. KorSTS 데이터셋에 대하여, KLUE-RoBERTa를 SimCSE로 학습한 모델들 (비지도, 지도, 멀티 태스크)의 훈련 전후 토큰들의 Self-Similarity의 변화.

표 2, 3은 KorSTS 데이터셋에서 각각 BERT와 RoBERTa의 대조학습된 모델에 대한 SSC 측정 결과이다. 빈도수가 많은 토큰들에 대하여, self-similarity가 가장 크게 변한 토큰들은 SS(↓)로, 가장 크게 변한 토큰들은 SS(↑)로 구분하였다. 전자의 경우 대조 학습 후 가장 문맥화가 많이 된 토큰들이며, 후자의 경우 반대로 대조학습의 결과 문맥화가 가장 덜 된 토큰들을 의미한다. 참고문헌 [17]의 아이디어를 따라서, 후자의 토큰들은 대조학습 중 의미 공간의 임베딩을 만드는 데 중요한 역할을 할 가능성이 높다고 가정한다.

표 2, 3의 결과를 통해 흥미로운 사실들을 몇 가지 발견할 수 있었다. 먼저, 이방성 문제가 해결되지 않은 모델일수록 한국어와 직접적 관련이 없는 토큰이나 특수 토큰 등 의미론적으로 쓸모 없는 것들이 주로 등장하는 점을 확인하였다. 그 예로, 비지도 대조학습으로 학습한

모델이 이방성 문제를 해결하는 데 가장 어려움을 겪기 때문에, 숫자나 기호 특수 토큰이나 의미 없는 토큰들이 등장하는 것을 확인할 수 있다. 다음으로, 참고문헌 [17]에서의 영어 기반 언어 모델 분석 결과와 유사하게, 문맥화가 되지 않은 토큰들 (SSC가 큰 토큰들)이 대체적으로 명사인 것을 확인하였다. 특히, 의미론적으로 문맥과 관련이 없을 가능성이 큰 요일 (수요일, 목요일, 금요일)이나 고유명사 (오바마, 우크라이나)가 등장한 점은 굉장히 흥미롭다. 또한 SSC가 작은, 즉 문맥화가 많이 된 토큰들의 경우 조사, 보조 용언, 의존 명사 등 문장의 실질적 문맥을 결정하지 않는 품사 혹은 문법 요소임을 볼 수 있었다.

표 4는 KorNLI 데이터셋에 BERT의 대조 학습 모델에 대하여 SSC를 측정한 결과이다.

비지도 대조학습		지도 대조학습		멀티 태스크	
SS (↓)	SS (↑)	SS (↓)	SS (↑)	SS (↓)	SS (↑)
0 ##됨	-	##하	제안	##으	제안
1 .	##마	##한다	보통	##될	디
2 [CLS]	V	##으	영어	##합니다	대답
3 물론	##벼	##합니다	대답	##합니다	영어
4 ##십시오	디	##이	디	##하	보통
5 그렇	비	##을	특별	##를	특별
6 ##한다	날	##를	머리	##을	대위
7 [SEP]	U	##기	웹	##할	머리
8 ##합니다	뉴	##될	대위	##이	농담
9 ##될	6	##으로	뉴스	##하여	웹

표 4. KorNLI 데이터셋에 대하여, KLUE-BERT를 SimCSE로 학습한 모델들 (비지도, 지도, 멀티 태스크)의 훈련 전후 토큰들의 Self-Similarity의 변화.

KorNLI의 경우에도 KorSTS와 유사한 형태의 결과를 얻을 수 있었다. 다만, KorSTS와 지도 대조학습과 멀티 태스크의 SS(↑) 결과의 차이점이 존재하였는데, 앞선 경우 고유 명사나 특수한 명사들이 등장했다는 점과 달리 KorNLI에서는 일반적으로 볼 수 있는 명사들이 등장하였다. 아마도 KorNLI가 추론을 요구하는 데이터셋이라, 이러한 특성이 고려되어 문맥에 따라 그 의미가 바뀌지 않을 토큰들 (제안, 대답 등)이 뽑힌 것으로 보인다.

4. 실험 및 분석 결과

본 장에서는 두 가지의 실험을 진행하였다. 먼저, 앞서 소개하였던 지표들을 대조학습이 진행된 한국어 언어 모델에서 측정하여 학습 전후의 차이를 보았다. 여기서 얻은 결과와 예비 연구의 SSC 실험을 통해 얻은 결과를 바탕으로 한국어의 대조학습 기반 문장 임베딩의 하위 문맥화를 일으키는 토큰들의 후보를 정리한다. 이 때, 한국어의 언어학적 특성을 고려하여 해당 토큰들이 실제 어휘의 특성과 부합하여 하위 문맥화에 기여함을 보인다. 두 번째 실험에서는, 만들어진 토큰 후보들을 기반으로 하여, 대조학습 문장 임베딩 모델의 코사인 유사도를 측정하는 간단한 실험으로 증강 방법의 유효성을 입증하며, 향후 벤치마크 및 데이터셋 설계와 강건한 훈련 방법론

에 밑거름이 되길 바란다.

4.1 지표 분석 실험 결과

Intra-sentence similarity의 경우 참고문헌 [17]에서 보고된 결과와 유사하게 마지막 출력 층 (12번 층)에서 급격히 상승하는 것을 볼 수 있었다. 그림 1, 2에서 볼 수 있는 것처럼, 모든 모델에서 같은 결과를 확인할 수 있다. 사전 학습 모델에서는 이러한 경향성이 없었기 때문에, 대조학습의 결과로 생긴 성능 향상의 주요 요인으로 보인다.

앞서 지표에 대한 설명처럼, 높은 intra-sentence similarity는 의미 공간의 활용도를 평가하는 지표로 활용될 수 있다. 이 지표가 높다는 것은 문장 임베딩이 하나의 중요 토큰에 수렴되고 있을 가능성을 나타낸다. 이 때 중요 토큰은 앞서 사전 실험에서 확인한 훈련 전후 문맥화가 덜 일어난 (SS↑) 토큰들일 가능성이 높다. 그 이유는, 문맥에 따라 의미가 변화하지 않기 때문에 대조학습의 특성을 고려했을 때 의미 공간에서 문장 표현의 기준 임베딩으로 활용될 수 있기 때문이다.

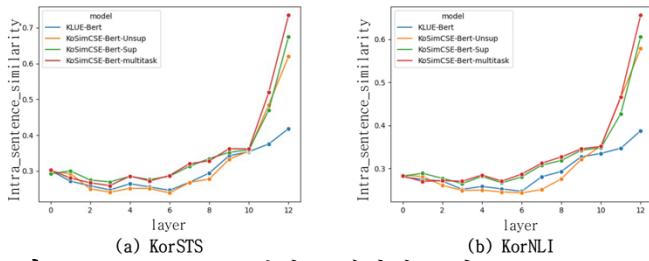


그림 1. BERT 모델의 레이어 별 Intra-sentence Similarity 측정 결과.

다음으로는, 토큰 빈도수를 기준으로 하여 SSC의 결과를 확인하였다. 그림 3의 경우 x축으로 빈도수 별로 토큰을 나열하고, y축으로 이들의 SSC를 시각화하였다. 빈도수가 가장 높은 토큰들은 일반적으로 [CLS], [SEP] 등의 특수 토큰들이다. 그림 3에서 알 수 있듯이, 이들 특수 토큰들의 경우 SSC가 작게 변하기 때문에 (음수인 경우도 존재), 문맥화가 잘 일어났다 보여진다. 이는 [CLS]나 [SEP], [PAD]의 역할이 문장 임베딩을 대표하거나, 혹은 최대 길이를 맞추기 위한 패딩(padding)을 만든 것에 있음을 고려했을 때 적절한 결과로 보인다. 특히, 이전 연구들에서 지적되었던 빈도수 편향 [16, 17]을 한국어 언어 모델에서도 확인할 수 있었다. 빈도수와 SSC가 반비례하는 경향이 있음을 볼 수 있었으며, 빈도수가 높을 수록 문맥화가 많이 일어나는 것을 확인할 수 있었다. 즉, 의미론적으로 문맥과 상관없이 그 뜻이 일정해야 하는 단어가 허위 문맥화가 되고 있었다.

이는 표 2, 3, 4의 결과에서도 확인할 수 있는데, ‘그러나, 현재, 또한, 물론’ 등의 단어들은 문맥에 따라 같은 의미를 가져야한 것이 자명하는데 불구하고, 음의 SSC가 큰 것을 확인할 수 있다. 참고문헌 [17]의 결과에서는 주로 불용어나 영어 언어상의 기능어들이 이러한 특성을 보였다. 이와 비슷한 관점에서, 본 논문은 아래의 4.2절을 통해 한국어 언어 모델에서의 허위 문맥화

토큰을 나누는 기준을 제안하고자 한다.

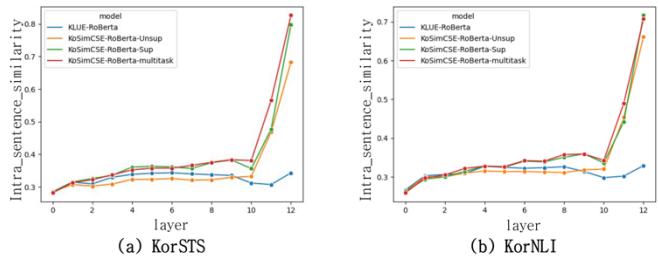


그림 2. RoBERTa 모델의 레이어 별 Intra-sentence Similarity 측정 결과.

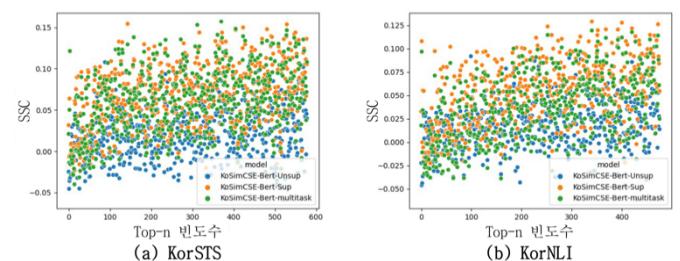


그림 3. BERT 모델의 토큰 빈도수별 SSC 결과.

4.2 한국어 언어 모델의 허위 문맥화 토큰

허위 문맥화 토큰을 나누는 기준을 제안하기 위하여 한국어와 영어의 언어학적 차이를 고려할 필요가 있다. 기존 한국어의 의존 해석을 위한 형태-통사적 품사 분류 체계와 관련된 연구 [1]가 있었으며 이는 전통적인 통사적 체계를 따른다. 영어는 내용어와 기능어로 구분되어 문맥화와 관련된 개념이 명확한 것과 달리 한국어 구성 성분은 내용어와 기능어의 결합 형태로 구성되며, 각 성분 간의 의미 있는 의존관계가 존재한다 [2]. 또한, 주술관계인 영어문장과는 달리 한국어는 화제-평언의 관계로써 동사가 문미에 위치하기 때문에 [4], 동사가 문장의 의미를 결정하는 중요 요소로 활용된다. 이와 달리 기존 연구 [17]에서는 영어의 명사, 형용사 등에 초점을 맞추었다.

이외에도 품사 분류와 관련하여 여러 방안과 연구가 수행되었다 [5, 6, 7]. 본 논문에서는 널리 사용되는 전통적인 품사 분류 체계를 따르되 앞서 언급한 영어와 한국어의 언어학적 차이점에 주목하고자 한다. 특히 영어와 달리 한국어는 교착어이기 때문에 형태소 단위의 토큰들이 문법적 역할을 결정할 가능성이 높다. 해당 토큰들은 대부분 문맥화가 일어나는 것이 일반적으로 보여지며, 대조학습의 결과가 이와 상응함을 표 1, 2, 3에서 확인할 수 있었다. 영어 언어 모델에서는 불용어들이 기능어로 많이 등장한 것과는 달리, 한국어 언어 모델의 경우 그렇지 않은 토큰들도 존재함을 볼 수 있었다.

본 논문에서는 허위 문맥화를 해결하는 데 기여하는 토큰들을 문맥 중심어라 명명하고, 그렇지 않은 토큰들을 문맥 기능어로 분류한다. 표 7과 같이, 문맥 중심어에는 영어와 유사하게 명사와 형용사 용언 뿐만 아니라, 숫자, 고유명사 등이 더 사용되었다. 문맥 기능어의 경우 기존 영어 언어 모델과 차이를 가지는데, 조사 뿐만

방법	문장 1	문장 2	유사도
원본	한 여성이 다른 여성의 발목을 재고 있다	한 여자는 다른 여자의 발목을 측정한다	79.85
원본	한 여성이 다른 여성의 발목을 재고 있다	한 남자가 노래를 부르며 기타를 연주하고 있다	42.54
문맥 기능어 제거	한 여성이 다른 여성의 발목을 재고 있다	한 여자는 다른 여자의 발목을 측정한다	85.02 (↑)
문맥 기능어 제거	한 여성이 다른 여성의 발목을 재고 있다	한 남자가 노래를 부르며 기타를 연주하고 있다.	42.84 (↑)
문맥 기능어 추가	한 여성이 다른 여성의 발목을 재고 있다 한다	한 여자는 다른 여자의 발목을 측정한다 한다	77.30 (↓)
문맥 기능어 추가	한 여성이 다른 여성의 발목을 재고 있다 한다	한 남자가 노래를 부르며 기타를 연주하고 있다 한다	42.15 (↓)
문맥 중심어 추가	한 소 여성이 다른 여성의 발목을 재고 있다	한 여자는 다른 여자의 발목을 측정한다	76.82 (↓)
문맥 중심어 추가	한 소 여성이 다른 여성의 발목을 재고 있다	한 소 남자가 노래를 부르며 기타를 연주하고 있다	53.45 (↑)
문맥 중심어 반복	한 소(x5) 여성이 다른 여성의 발목을 재고 있다	한 여자는 다른 여자의 발목을 측정한다	68.43 (↓)
문맥 중심어 반복	한 소(x5) 여성이 다른 여성의 발목을 재고 있다	한 소(x5) 남자가 노래를 부르며 기타를 연주하고 있다	64.91 (↑)

표 5. KorSTS 테스트 데이터셋 샘플에서의 적대적 데이터 증강에 따른 유사도 변화. 비지도 대조학습한 BERT 모델이 사용되었다.

방법	문장 1	문장 2	유사도
원본	그녀는 아직 그 안에 있었다	그녀는 여전히 근처에 있었다	81.14
원본	그녀는 아직 그 안에 있었다	그녀는 흔적도 없이 사라졌어요	51.47
문맥 기능어 제거	그녀는 아직 그 안에 있었다	그녀는 여전히 근처에 있었다	81.39 (↑)
문맥 기능어 제거	그녀는 아직 그 안에 있었다	그녀는 흔적도 없이 사라졌어요	47.51 (↓)
문맥 기능어 추가	그녀는 아직 그 안에 있었다고	그녀는 여전히 근처에 있었다고	79.92 (↓)
문맥 기능어 추가	그녀는 아직 그 안에 있었다고	그녀는 흔적도 없이 사라졌고	51.45 (↓)
문맥 중심어 추가	그녀는 아직 제안 그 안에 있었다	그녀는 여전히 근처에 있었다	76.01 (↓)
문맥 중심어 추가	그녀는 아직 제안 그 안에 있었다	그녀는 제안 흔적도 없이 사라졌어요	54.37 (↑)
문맥 중심어 반복	그녀는 아직 제안(x5) 그 안에 있었다	그녀는 여전히 근처에 있었다	74.85 (↓)
문맥 중심어 반복	그녀는 아직 제안(x5) 그 안에 있었다	그녀는 제안(x5) 흔적도 없이 사라졌어요	58.32 (↑)

표 6. KorNLI 검증 데이터셋 샘플에서의 적대적 데이터 증강에 따른 유사도 변화. 지도 대조학습한 BERT 모델이 사용되었다.

아니라 보조 용언, 종결 어미, 의존 명사 등도 고려되었다. 나아가, 접속사나 부사들도 문맥 기능어로 사용될 수 있다.

4.3 문장 임베딩에 대한 적대적 데이터 증강

앞서 제안한 문맥 중심어와 기능어를 활용하여, 간단한 적대적 데이터 증강 방법론들을 제안한다. 실험은 대조학습된 한국어 언어 모델을 사용하였으며, KorSTS 테스트 데이터셋의 샘플과 KorNLI 검증 데이터셋의 샘플을 사용하였다. KorSTS의 경우 유사도가 높은 문장 쌍을 고르고, 이와 관련 없는 문장을 추가로 선택하였다. KorNLI의 경우 문장 하나를 기준으로 entailment와 contradiction 문장을 선택하여 실험을 진행하였다. 제안하는 방법론들의 유효성을 검증하기 위하여 적대적 데이터 증강이 적용되기 전후의 코사인 유사도를 측정하였다.

토론 종류	종류	예시
문맥 중심어	숫자	12, 13
	고유명사	오바마, 수요일
	명사	폭발, 석방
	형용사 용언	시끄러운, 차갑
문맥 기능어	보조 용언	##하다
	서술격 조사	##이다
	종결 어미	##어요
	조사	##을, ##를, ##이
	의존 명사	명
	접속사	그러나, 또한
	부사	그러나, 물론
	기타 토큰	기호, 특수 토큰

표 7. 한국어 언어 임베딩의 허위 문맥화와 관련된 문맥 중심어와 문맥 기능어의 구분.

표 5, 6은 적대적 데이터 증강 방법론과 문장 쌍, 그리고 문장 쌍 사이의 유사도를 기록한 결과이다. 본 논문은 문맥 기능어가 문장 유사도에 악영향을 끼칠 것을 가정하여 문맥 기능어의 추가, 제거 실험을 진행하였다. 그 결과, 문맥 기능어를 제거하였을 때 유사도가 증가하고, 추가하였을 때 감소함을 확인할 수 있었다. 다음으로 문맥 중심어가 문장의 문맥을 결정하는 데 크게 기여한다 판단하여, 이를 추가하거나 반복하는 실험을 진행하였다. 문맥 중심어를 기본 문장에 추가 또는 반복하였을 때, 의미론적으로 유사한 문장의 경우 유사도가 감소하는 것을 볼 수 있다. 문맥 중심어를 기본 문장과 의미론적으로 다른 문장에 모두 추가하였을 경우에는, 해당 단어로 인해 유사도가 증가하는 것을 확인할 수 있었다.

5. 결론

영어 및 한국어 기반의 사전 학습 언어 모델이 큰 발전을 이루고, 이에 따라 질 좋은 문장 표현의 생성이 가능해지고 실생활의 다양한 작업들에 널리 활용되고 있음은 분명하다. 대조학습 등 사전 학습 언어 모델에 널리 사용되는 방법론을 분석하고, 나아가 허위 문맥화를 파악하는 실험은 기술의 강건성에 도움이 될 수 있기 때문에, 본 논문은 이전 영어 언어 모델에서의 연구를 차용하여 분석 실험을 진행하였다. 나아가, 한국어의 언어학적 특성을 고려하여 허위 문맥화와 관련된 토큰들을 분류하는 기준을 제안하였으며, 이들을 활용하여 적대적 데이터 증강이 가능함을 간단한 실험으로 확인하였다. 향후 이를 바탕으로 새로운 벤치마크 제작 등으로 확장하고자 한다.

감사의 글

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No.2020-0-01373, 인공지능대학원지원(한양대학교))을 받아 수행되었습니다.

참고문헌

- [1] 홍영국, 이종혁. “한국어 의존 해석을 위한 형태 - 통사적 품사 분류 체계.” *정보과학회논문지(B)*, 22(9), 1375-1383, 1995.
- [2] 최선화, 박혁로. “한국어 확률 의존문법 학습.” *한국정보과학회 학술발표논문집*, 30(1B), 513-515, 2003.
- [3] Bohm, David. *Wholeness and the implicate order*. Routledge, 2005.
- [4] 염행일. *Strategies in Consecutive Interpretation - With Special Emphasis on Interpretation from Korean into English*. 통번역교육연구, 6(2), 165-181, 2008.
- [5] 구본관(Koo, Bon-Kwan), “국어 품사 분류와 관련한 몇 가지 문제.” *형태론*, 12권, 2호, 179-199, 2010.
- [6] 최형용. “유형론적 관점에서 본 한국어의 품사 분류 기준에 대하여-분류 기준으로서의 ‘형식’을 중심으로.” *형태론*, 14(2), 233-263, 2012.
- [7] 목정수. “품사론, 형태에서 의미로 아니면 기능에서 형태로?” *한국언어학회 학술대회지*, 17-43, 2015.
- [8] Vaswani, A. et al. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [9] Cer, D. et al. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.
- [10] Devlin, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [11] Liu, Y. et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [12] Ethayarajh, K. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 55-65, 2019.
- [13] Ham, J. et al. KorNLI and KorSTS: New Benchmark Datasets for Korean Natural Language Understanding. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, p. 422-430, 2020.
- [14] Park, S. et al. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*, 2021.
- [15] Gao, T. et al. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 6894-6910, 2021.
- [16] Rajaei, S. et al. How Does Fine-tuning Affect the Geometry of Embedding Space: A Case Study on Isotropy. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*, p. 3042-3049, 2021.
- [17] Xiao, C. et al. On Isotropy, Contextualization and Learning Dynamics of Contrastive-based Sentence Representation Learning. In: *Findings of the Association for Computational Linguistics: ACL 2023*, p. 12266-12283, 2023.