BERT 레이어에 따른 동형이의어 의미 표현 비교

강일민○, 최용석, 이공주
충남대학교
ilmin0325@naver.com, yseokchoi@cnu.ac.kr, kjoolee@cnu.ac.kr

Comparison of Homograph Meaning Representation according to BERT's layers

Il Min Kang○, Yong-Seok Choi, Kong Joo Lee
Chungnam National University

요 약

본 논문은 BERT 모델을 이용하여 동형이의어의 단어 표현(Word Representation) 차이에 대한 실험을 한다. BERT 모델은 Transformer 모델의 인코더 부분을 사용하여 양방향을 고려한 단어 예측과 문장 수준의 이해를 얻을 수 있는 모델이다. 실험은 동형이의어에 해당되는 단어의 임베딩으로 군집화를 수행하고 이를 Purity와 NMI 점수로 계산하였다. 또한 각 단어 임베딩 사이를 코사인거리(Cosine Distance)로 계산하고 t-SNE를 통해 계층에 따른 변화를 시각화하였다. 군집된 결과는 모델의 중간 계층에서 점수가 가장 높았으며, 코사인거리는 8계층까지는 증가하고 11계층에서 급격히 값이 변하는 것을 확인할 수 있었다.

주제어: 동형이의어, BERT, 군집화, Purity, Normalized Mutual Information, 코사인거리

1. 서론

동형이의어는 형태가 같으면서 동시에 의미가 다른 단어들을 뜻한다. 단어의 형태는 일치하지만 문장 내에서 서로 다른 의미를 가지고 있기 때문에 주변 문맥을 통해 구분한다. 표 1은 동형이의어에 대한 예시이다.

표 1. 동형이의어 예시

(1) 물가는 오르고 임금은 물가 인상을 따르지 못하니 생활이 어렵다.
(2) 세종은 조선 4대 임금이다.

표 1-(1)에서 '임금'의 의미는 '근로자가 노동의 대가로 사용자에게 받는 보수'를 뜻한다. 표 1-(2)에서는 '군주 국가에서 나라를 다스리는 우두머리'를 뜻한다. 이처럼 동형이의어는 상이한 뜻을 가지고 있어 구분이 필요하다. 문장 내에서 단어의 의미가 중요시되는 과제에서는 결과에 큰 영향을 미칠 수 있다.

Word2vec[1]와 같은 기존의 정적 단어 표현(Static Word Representation)에서는 단어마다 고정된 임베딩을 사용한다. 모든 문맥에 대하여 동일한 단어 표현을 사용하기 때문에 동형이의어에 대한 표현을 고려하지 않는다. 반면 문맥 의존적 단어 표현(Contextual Word Representation)은 단어마다 고정된 벡터가 아닌 문장의 문맥에 따라서 단어의 임베딩이 달라지는 방식이다. ELMo[2], GPT[3], BERT[4]가 이에 속한다. 위 모델들은 모두 문맥 정보를 반영하고 기존의 단어 표현보다 더 나은 성능을 보여주었다.

본 연구는 BERT[4] 모델을 이용하여 동형이의어의 표현 차이에 대한 실험을 진행한다. 또한 BERT[4]의 각 계층에서 단어 임베딩이 어떤 변화가 생기는지에 대한 실험을 수행한다.

2. 관련 연구

BERT[4]는 Transformer[5] 모델의 인코더 부분을 사용한 모델이다. BERT는 양방향을 고려한 단어 예측과 문장 수준의 이해를 얻는 것을 목표로 한다. 따라서 일반적인 언어 모델을 사용하는 대신 언어 모델의 마스킹(Mask Language Model)과 다음 문장 여부를 예측(Next Sentence Prediction)하는 작업으로 사전 훈련을 시킨다.

[6]의 연구는 영어권에서 BERT[4]의 학습된 표현들을 기반으로 실험을 진행하였다. 실험은 CoNLL 2000 데이터셋에서 구문 라벨을 이용하여 군집화를 진행하고 이를 NMI(Normalized Mutual Information)로 평가하였다. 또한, 각 층에서 학습된 표현들을 이용하여 층마다 어떤 언어적 특징을 가지고 있는지를 비교해보았다. 구문 라벨을 이용한 군집화의 실험 결과는 1 계층에서 가장 좋은 점수를 얻었다. 언어적 특징을 비교해보기 위한 실험에서는 언어 표면(surface)적의 표현은 낮은 계층, 언어 구문(syntactical)적 표현은 중간 계층 그리고 언어 의미(semantical)적 표현은 상위 계층에서 특징을 포착해내고 있음을 보여주었다.

본 연구에서는 동형이의어가 BERT[4] 모델에서 어떻게 표현되고 있는지에 대해 비교해보고자 한다. 또한 각 계층에서 동형이의어 단어들이 어떤 변화가 생기는지에 대한 실험을 수행해보고자 한다.

3. 실험 데이터 수집

실험 데이터는 형태소, 품사, 동형이의어 부착 말뭉치[1][7]에서 추출하였다. 문장의 해당 단어는 국립국어원 표준국어대사전[2]의 단어 의미 번호가 부여되어 있다. 수집한 동형이의어는 총 499개이고, 1,225개의 의미를 가지고 있다. 각 동형이의어 마다 2개 이상의 의미가 있으며 평균 2.45이개다.

총 문장 수는 65513개이며, 각 동형이의어 당 평균 문장 수는 131.29개, 각 의미 당 문장 수는 53.48개이다. 문장 당 평균 어절 수는 14.35개이며, BERT[4]의 기본 토큰 단위인 WordPiece[8]로 나누었을 경우 각 문장마다 평균 30.29 토큰이다.

표 2. 실험 데이터의 의미 번호와 의미 예시

단어/의미 번호	의미
사원/02	「명사」 종교의 교당을 통틀어 이르는 말
사원/04	「명사」 회사에서 근무하는 사람
수용하다/05	「동사 어떠한 것을 받아들이다
수용하다/03	「동사」 범법자, 포로, 난민, 관객, 물품 따위를 일정한 장소나 시설에 모아 넣다

표 3. 실험 데이터

단어 개수	499
의미 개수	1,225
총 문장 수	65,513
단어 당 평균 문장 수	131.29
의미 당 평균 문장 수	53.48
문장 당 평균 어절 수	14.35
문장 당 평균 토큰 수	30.29

실험을 위한 데이터에 포함된 동의어 집합은 네이버 동의어 사전을 통해 수집했다. 동의어 집합의 개수는 100개이며 총 261개의 의미가 있다. 다음 표 4는 수집한 동의어 집합에 대한 예시이다.

표 4. 동의어 실험 데이터 예시

단어/의미 번호	의미
수련/06	「명사」 인격, 기술, 학문 따위를 닦아서 단련함
수행/01	「명사」 행실, 학문, 기예 따위를 닦음
사원/02	「명사」 종교의 교당을 통틀어 이르는 말
사찰/02	「명사」 승려가 불상을 모시고 불도(佛道)를 닦으며 교법을 펴는 집
절/01	「명사」 승려가 불상을 모시고 불도(佛道)를 닦으며 교법을 펴는 집
조류/01	「명사」 조강의 척추동물을 일상적으로 통틀어 이르는 말
새/03	「명사」 몸에 깃털이 있고 다리가 둘이며, 하늘을 자유로이 날 수 있는 짐승을 통틀어 이르는 말

4. 실험 환경

BERT[4] 모델은 ETRI에서 제공한 한국어 BERT 언어모델[3]을 사용한다. BERT의 모델은 12개의 계층과 768개의 차원의 은닉 계층으로 구성되어 있다. 주의(Attention)의 헤드 개수는 12개로 되어 있다.

BERT[4]는 토큰의 기본 단위로 형태소 단위의 WordPiece[8]을 사용한다. 형태소 분석된 문장을 입력으로 넣었을 때 BERT의 토큰 단위로 나눠지게 된다. 이때 한 단어가 BERT의 어휘 집합(Vocabulary)에 없을 경우 여러 개의 토큰으로 나눠지는 현상이 발생한다. 나눠진 토큰이 포함된 문장은 실험 데이터 중 3,471문장이며 이는 전체 문장의 5.3%이다. 실험에는 여러 개로 나눠진 토큰에 대한 임베딩은 각 토큰의 평균을 구하여 한 단어의 임베딩으로 표현하였다. 표 5는 BERT WordPiece 단위에 대한 예시이다.

표 5. WordPiece단위 예시

단어	WordPiece
(1) 보이/VV	보이/VV
(2) 임금/NNG	임금/NNG
(3) 단정하/VV	단+정+하/VV
	단+정하/VV

5. 실험 방법 및 평가

5.1 실험 방법

본 연구에서는 계층적 군집(Hierarchical Clustering)[9] 중 Bottom-up 방법으로 병합 군집(Agglomerative Clustering)을 사용한다.

병합 군집 알고리즘은 시작할 때 각 포인트를 하나의 클러스터로 지정하고 종료 조건으로 지정된 클러스터 개수를 만족할 때까지 가장 비슷한 두 클러스터를 합쳐나가는 방법이다. 실험은 동형이의어에 해당되는 단어의 임베딩으로 의미의 개수만큼 군집화를 한 결과와 전체 데이터의 임베딩을 가지고 총 의미의 개수(1,225개)로 군집화를 한 결과를 사용한다.

병합 군집 알고리즘은 scikit-learn[4]을 사용했다.

5.2 평가 방법

본 연구에서는 군집화 결과를 평가하는 방법으로 Purity와 NMI(Normalized Mutual Information)를 사용한다.

군집화 실험은 2가지로 나누어 진행한다. 첫 째, 동형이의어에 해당되는 단어의 의미 개수만큼을 군집화를 각각 수행하여 평균값을 구한다(Purity/NMI 단어). 둘 째, 동형이의어 단어의 구분 없이 실험 데이터에 포함된 전체 의미 개수만큼을 군집화를 수행하여 점수를 계산한다(Purity/NMI 전체).

BERT[4] 모델에서 각 계층의 단어 임베딩에 따른 변화를 보기 위해 코사인 거리(Cosine distance)를 계산해본다.

1) http://nlplab.ulsan.ac.kr/doku.php
2) https://stdict.korean.go.kr/main/main.do
3) http://aiopen.etri.re.kr/service_dataset.php
4)https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

표 7. 계층 별 군집 Purity, NMI 계산 결과

	Layer1	Layer2	Layer3	Layer4	Layer5	Layer6	Layer7	Layer8	Layer9	Layer10	Layer11	Layer12
Purity 단어	0.7975	0.845	0.8725	0.89	0.8925	**0.8975**	0.895	0.895	0.895	0.8925	**0.8975**	0.895
Purity 전체	0.8410	0.8779	0.9003	0.9097	**0.9098**	0.9089	0.8995	0.8840	0.8680	0.8433	0.8452	0.8446
NMI 단어	0.4698	0.5804	0.6525	0.6765	0.6850	**0.6886**	0.6870	0.6755	0.6688	0.6605	0.6712	0.6665
NMI 전체	0.9336	0.9500	0.9560	**0.9584**	0.9581	0.9568	0.9511	0.9417	0.9319	0.9196	0.9197	0.9179

표 8. 계층 별 코사인거리 계산 결과

	Layer1	Layer2	Layer3	Layer4	Layer5	Layer6	Layer7	Layer8	Layer9	Layer10	Layer11	Layer12
CosDist-1	0.1233	0.1716	0.2057	0.2278	0.2443	0.2587	0.2783	0.3186	0.3130	0.2600	0.2016	0.2728
CosDist-2	0.3068	0.3608	0.4097	0.4488	0.4659	0.4835	0.4907	0.5302	0.5092	0.4096	0.3163	0.4265
CosDist-3	0.7930	0.7814	0.7463	0.7161	0.6746	0.6493	0.6148	0.6230	0.5738	0.4437	0.3405	0.4725

5.2.1 Purity

본 연구에서는 실험 데이터를 병합 군집을 통해 분류하여 군집의 Purity를 계산한다(수식 1).

총 문장 수(N)이며 분류된 군집(w_k)이고 정답 클래스(c_j)이라 할 때, Purity는 빈도수가 가장 높은 정답 클래스를 분류된 군집에 대표 값으로 할당하여 계산하는 방법이다.

$$Purity(W, C) = \frac{1}{N}\sum_{k}\max_{j}|w_k \cap c_j| \quad (수식 1)$$

5.2.2 NMI(Normalized Mutual Information)

Purity는 유사성이 높은 개체의 군집된 결과를 이용하여 계산하기 때문에 실제 정답과 비교 평가에는 부적합하다. 이를 보안하기 위해 본 연구에서는 Purity 이외에 다른 평가 척도를 도입했다.

Mutual Information[10]란 두 확률 변수간의 상호 의존도를 나타내는 지표이다. 이를 0과 1사이 값이 되도록 정규화한 지표가 NMI이다. 이를 이용해 실제 정답과 군집된 결과의 상호 의존도를 계산한다.

수식 2는 NMI에 대한 수식이다. H는 엔트로피(Entropy) 함수이며, I는 W와 C사이의 Mutual Information이다.

$$NMI(W, C) = \frac{2 \times I(W;C)}{[H(W) + H(C)]} \quad (수식 2)$$

5.2.3 코사인거리(Cosine Distance)

본 논문에서는 BERT[4] 모델의 각 계층에 따른 단어 표현의 변화에 대한 실험을 했다. 모델의 낮은 계층에는 단어만의 의미를 반영하고 계층이 높아질수록 단어의 의미 이외의 문장의 맥락에서 다른 단어와의 추가적인 정보가 반영되어 낮은 계층과는 다른 표현을 가질 것이라고 가정한다. 이를 증명하기 위해 동형이의어의 임베딩 사이의 유사도와 각 계층에 따른 변화에 대한 분석을 위해 단어 임베딩 사이의 코사인거리(Cosine Distance)를 측정했다. 낮은 계층에서는 단어의 의미만 반영되어 서로의 거리가 가까웠다면, 높은 계층에서는 의미 이외의 추가적인 정보로 인해 서로의 거리가 조금 멀어지고 동의어 사이 거리는 가까워져야 한다. 실험은 3가지 방법으로 진행한다(표 6).

표 6-(1)은 의미 번호가 동일한 단어 임베딩 사이의 코사인거리를 측정했다(CosDist-1). 표 6-(2)은 형태는

표 6. 코사인거리 실험 방법과 예제

동형이의어 관계 : 사원/02, 사원/04 동의어 관계 : 사원/02, 사찰/02	
(1)	같은 형태, 같은 의미인 단어 임베딩 사이의 거리
	사원/02의 각 단어의 임베딩 사이의 코사인 거리
(2)	같은 형태, 다른 의미인 단어 임베딩 사이의 거리
	사원/02, 사원/04 사이의 코사인 거리
(3)	다른 형태, 같은 의미인 단어 임베딩 사이의 거리
	사원/02, 사찰/02 사이의 거리

같지만 의미가 다른 즉 동형이의어 관계인 단어의 임베딩간의 코사인거리를 계산했다(CosDist-2). 표 6-(3)은 실험 데이터에 존재하는 다른 형태 같은 의미 즉 동의어 관계 집합을 활용하여 임베딩 사이의 거리를 구하는 실험을 진행한다(CosDist-3). 또한 변화 추이를 t-SNE[11]를 이용하여 시각화했다.

6. 실험 결과

모델의 각 계층에서의 군집에 대한 실험 결과는 표 7에 나타내었다. 표 7의 각 동형이의어에 대한 Purity(Purity 단어)는 6과 11 계층에서 0.8975로 가장 높았고 1 계층에서 0.7975로 가장 낮았다. 동형이의어 구분 없이 의미에 대한 Purity(Purity 전체)는 5 계층에서 0.9097로 가장 높았으며, 1 계층에서 0.841로 가장 낮았다.

각 동형이의어의 NMI(NMI 단어)는 6 계층에서 0.6886로 가장 높았고 1 계층에서 0.47로 가장 낮았다. 동형이의어 구분 없이 의미에 대한 NMI(NMI 전체)는 4 계층에서 0.9584로 가장 높았고 12 계층에서 0.918로 가장 낮았다. 군집된 결과를 살펴보면 BERT[4]의 중간 계층에서 가장 높은 점수를 받은 것을 확인할 수 있었다.

표 8은 BERT의 각 계층에서 단어의 표현을 코사인거리로 실험한 결과이다. CosDist-1과 CosDist-2 실험은 모두 8 계층까지는 거리 값이 증가한다. 하지만 그 이후 감소하다 11 계층에서 급격한 변동이 있다. CosDist-3 실험은 계층이 높아질수록 거리 값이 낮아지지만 11 계

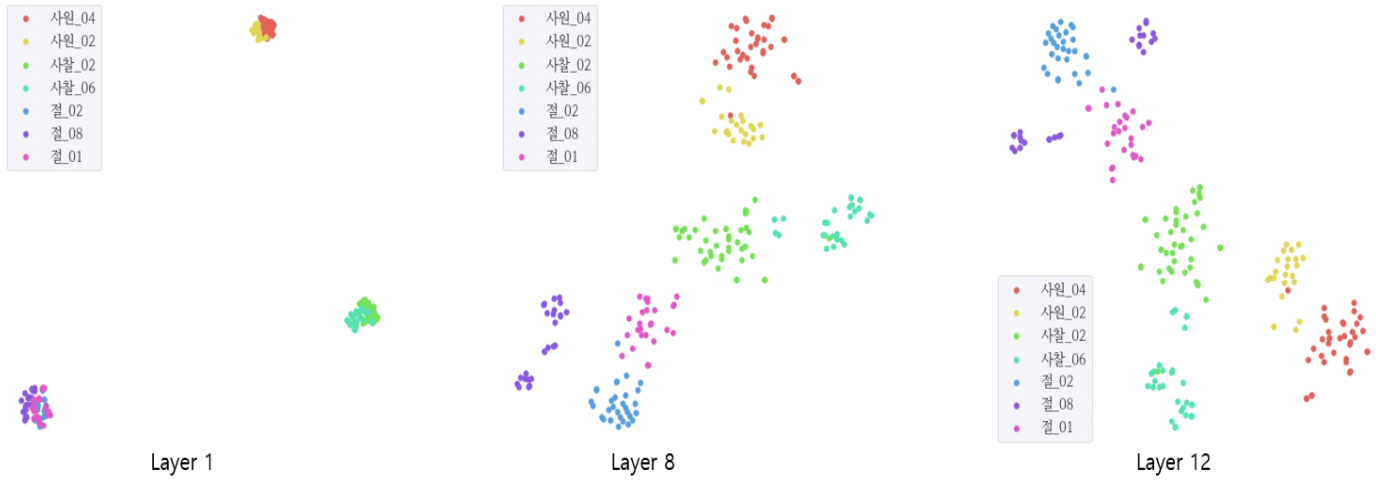

그림 1. t-SNE를 이용한 BERT 모델 계층에서의 '사원', '사찰', '절' 단어 임베딩 표현

층에서 감소폭이 컸다. 그렇지만 12 계층에서 10 계층과 비슷한 결과를 보였다.

그림 1은 1, 8, 12 계층에서 '사원', '사찰', '절'의 단어 임베딩을 t-SNE[11]으로 구현한 것이다.

'사원'의 동형이의어는 '사원/02'와 '사원/04'가 있다. '사찰'의 동형이의어는 '사찰/02'와 '사찰/06'이 있다. '절'의 동형이의어는 '절/02', '절/08', '절/01'이 있다. '사원/02'의 동의어는 사찰/02'와 '절/01'이다.

1 계층에서는 동형이의어 단어별로 뭉쳐있는 것을 볼 수 있다. 8 계층에서는 동형이의어의 단어들이 흩어진 모습을 확인할 수 있고, 동의어들이 가까워진 모습을 볼 수 있다. 12 계층에서는 8 계층과 유사한 모습을 볼 수 있다.

6. 결론

본 연구에서는 한국어 BERT 모델을 이용하여 동형이의어의 표현 차이에 대한 실험을 수행하였다. 실험을 위해 동형이의어를 군집화를 시도해 보았다. 군집화 실험 척도는 Purity와 NMI를 사용하였다. 실험결과는 BERT의 중간 계층에서 Purity와 NMI 점수가 높은 것을 확인하였다.

추가적으로 BERT 모델의 각 계층에서의 동형이의어와 동의어 임베딩 표현을 이용하여 코사인거리를 계산해 보았다. 실험결과는 같은 형태, 같은/다른 의미인 단어 임베딩 간의 거리는 8 계층까지는 거리가 멀어지는 것을 볼 수 있다. 또한 다른 동의어간의 거리는 11 계층까지 거리가 가까워지는 것을 볼 수 있다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2019R1F1A1053136)

참고문헌

[1] Mikolov, Tomax, et al. "Efficient estimation of word representations in vector space". *arXiv preprint arXiv:1301.3781*, 2013.

[2] Peters, Matthew E., et al. "Deep contextualized word representations." *arXiv preprint arXiv:1802.05365*, 2018.

[3] Radford, Alec, et al. "Improving language understanding by generative pre-training." *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[4] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*, 2018.

[5] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

[6] Jawahar, Ganesh, et al. "What does BERT learn about the structure of language?." *57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy*. 2019.

[7] 신준철, and 옥철영. "한국어 품사 및 동형이의어 태깅을 위한 단계별 전이모델." 정보과학회논문지: 소프트웨어 및 응용 39.11 (2012): 889-901.

[8] Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv preprint arXiv:1609.08144*, 2016.

[9] Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 2005. 321-352.

[10] (URL) https://en.wikipedia.org/wiki/Mutual_information

[11] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-sne." Journal of machine learning research 9.Nov, 2008.