한국어 단어 정의 벡터를 이용한 단어 의미 모호성 해소

박정연○, 이재성
충북대학교
{parkjeongyeon, jasonlee}@chungbuk.ac.kr

Word Sense Disambiguation Using Korean Word Definition Vectors

Jeong Yeon Park○, Jae Sung Lee
Chungbuk National University

요 약

기존 연구에 따르면, 시소러스의 계층적 관계를 기반으로 압축한 의미 어휘 태그를 단어 의미 모호성 해소에 사용할 경우, 그 성능이 향상되었다. 본 논문에서는 시소러스를 사용하지 않고, 국어 사전에 포함된 단어의 의미 정의를 군집화하여 압축된 의미 어휘 태그를 만드는 방법을 제안한다. 또, 이를 이용하여 효율적으로 단어 의미 모호성을 해소하는 BERT 기반의 딥러닝 모델을 제안한다. 한국어 세종 의미 부착 말뭉치로 실험한 결과, 제안한 방법의 성능이 F1 97.21%로 기존 방법의 성능 F1 95.58%보다 1.63%p 향상되었다.

주제어: 단어 의미 모호성 해소, 군집화, 딥러닝, 의미 어휘, 사전

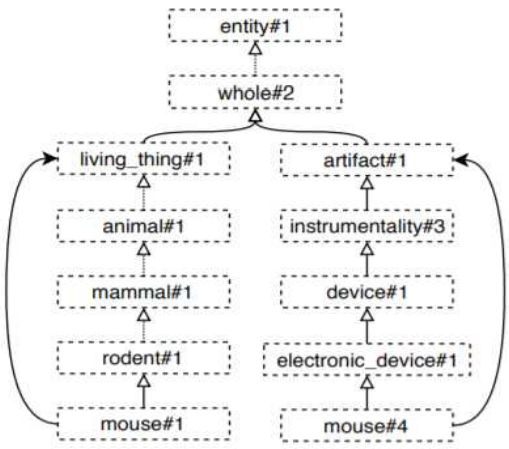
1. 서론

동형이의어는 문장내에서 문맥에 따라 다른 의미로 해석된다. 이와 같이 단어가 여러 가지 의미를 지님으로써 단어의 뜻을 쉽게 결정할 수 없는 경우 "단어 의미 모호성이 있다" 라고 한다. "단어 의미 모호성 해소"는 문장내에서 사용된 동형이의어의 정확한 의미를 결정하는 것을 말한다[1].

단어 의미 모호성 해소는 여러가지 방법으로 오랫동안 연구되어 왔으며, 최근 지도 학습 기반 방법이 많이 연구되고 있다. 그러나, 지도 학습 기반 방법은 학습 데이터의 부족으로 인해 희소 데이터를 처리해야 한다. 이러한 문제점을 해결하기 위해 시소러스 등의 지식을 사용하여 그 문제점을 보완하는 연구들이 발표되었다[2-5].

Huang et al.(2019)[2]는 사전 정의를 이용하는 GlossBERT를 제안하였다. 여기에서는 동형이의어가 포함된 문장과 동형이의어 각각의 의미에 해당되는 사전 정의(Gloss)를, BERT의 두 문장 비교 기능을 이용하여 직접 비교하여 의미를 결정하였다. Bevilacqua and Navigli(2020)[3]는 WordNet과 같은 관계 지식을 그래프로 인코드하여 이를 벡터로 계산할 수 있도록 했다. 즉, BERT를 활용하여 생성한 문장 임베딩과, WordNet의 유의어 집단과 이들 사이의 관계를 적절한 가중치로 고려하여 만든 벡터를 이용하여 신경망내에서 의미 예측 확률을 구하였다. Scarlini et al.(2020)[4]의 연구에서는 위키피디아의 의미 관련 페이지, BabelNet의 의미 어휘 정의, 의미 레이블이 붙은 학습데이터의 문맥(지도학습 방법의 경우)을 모두 활용한다. 즉, 앞의 3가지 문맥을 각각 BERT에 입력하여 벡터를 만들고 이를 종합하는 의미벡터를 만든 후, 테스트할 단어의 문맥벡터를 이 의미벡터들에 대한 1-nearest neighbor로 비교 선택하여 의미 모호성을 해소 하였다.

특히, Vial et al.(2019)[5]의 연구는 단어의 유의어, 상하위어, 반의어 등 여러 관계가 나타난 WordNet을 이용하여 의미 어휘 태그를 압축하고, 압축 태그를 seq labeling에 사용하여 의미 모호성을 효과적으로 해소하는 방법을 제안하였다. 의미 어휘 압축의 핵심은 시소러스의 계층 구조에서 동형이의어의 조상이 겹치지 않는 최대 상위 조상의 의미 태그를 공통의 태그로 만드는 것이다. 예를 들어 <그림 1>에서, 포유류 동물 쥐의 의미를 지닌 "mouse#1"과 컴퓨터 입력장치의 일종이라는 의미를 지닌 "mouse#4"에 대해, 공통의 조상 태그가 없도록 하면서 가질 수 있는 최상위 조상 의미 태그는 각각 "living_thing#1"과 "artifact#1"이다. 즉, "mouse#1", "rodent#1", "mammal#1", "animal#1" 대신 "living_thing#1"를 압축 의미 태그로 사용할 수 있고, "mouse#4", "electronic device#1", "device#1", "instrumentality#3"는 "artifact#1"를 압축 의미 태그로 대신 사용할 수 있다.



<그림 1> 상하위어 관계를 이용한 의미 어휘 태그 압축[5]

WordNet은 1985년부터 개발이 시작되어 현재도 개발이 지속적으로 이루어지고 있다. 그러나, 영어권 이외의 시소러스는 WordNet을 그대로 번역한 경우가 많으며, 자체 개발한 경우에도 그 규모가 WordNet에 비해 작거나 품질이 낮고, 공개적으로 쉽게 얻을 수 없는 경우가 많다[6].

본 논문에서는 WordNet과 같은 시소러스를 사용하지 않고 사전에 나타난 의미 어휘의 정의를 벡터화하고 군집화(clustering)하여 태그를 압축하는 방법을 제안한다. 또한 새로 구축한 의미 압축 태그를 단어 의미 모호성 해소에 사용할 수 있는 모델을 제안하고 이를 평가한다.

2. 모델

2.1 의미 어휘 군집화

본 연구에서는 사전에 기록된 의미 정의를 이용하여 의미 어휘를 압축한다. 우선, 모든 단어의 의미 정의는 Universal Sentence Encoder(USE)[7] 모델을 사용하여 의미 어휘 벡터로 표현한다. 다음으로 의미 어휘 벡터를 품사별로 구분하고, 동일한 품사 내에서 군집화를 진행한다. 이때, 군집화는 계층적 병합 군집화로 이루어지며, 병합 대상은 1. 벡터 유사도가 가장 높고, 2. 지정된 임계값(threshold)보다 유사도가 크며, 3. 병합하였을 때 그룹내에 동형이의어가 포함되지 않아야 한다.

계층적 병합 군집화는 n^2 크기의 유사도 행렬을 모두 검사하고, 군집을 최대 $n-1$번 합치는 연산이 수행되므로 $O(n^3)$의 시간 복잡도를 가진다. 또한, 의미 어휘의 벡터 유사도 거리 탐색을 위한 연산은 표준국어대사전의 434,192개 어휘가 포함되어 있고, 이를 수행하기 위해서는 약 702G 이상(= 434K * 434K * 4)의 유사도 행렬을 유지할 큰 메모리 공간을 필요로 한다. 이러한 시간 및 공간의 복잡성을 해결하기 위한 방법이 필요하다. 따라서 본 연구에서는 연산량이 비교적 적은 Euclidean distance 계산을 사용하여 연산 시간을 줄인다. (유사도는 Euclidean distance에 반비례함) 또한 적정한 크기의 메인 메모리 공간에서 군집화하기 위해 소규모 군집을 먼저 구한 후, 이를 대상으로 계층적 병합 군집화 방법을 적용한다.

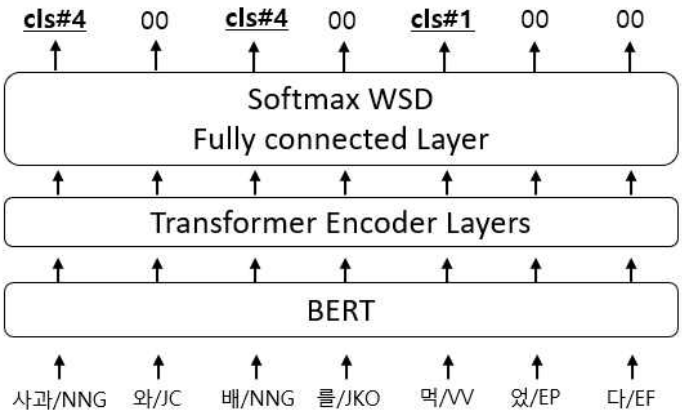
소규모 군집을 분리하기 위해서는, 주어진 K개의 군집으로 군집화하는 K-means 알고리즘[8]과 자동으로 적절한 군집 개수를 결정하는 Affinity Propagation(AP) 알고리즘[9]을 사용하여, 아래 조건에 따라 수행한다.

- AP를 동작할 수 있는 시스템 환경이라면 AP를 통해 초기 군집을 생성한다.
- 만약 AP가 동작할 수 없다면 시스템에서 AP를 동작시킬 수 있는 최대한의 데이터를 무작위 추출하고, 이 데이터에 대해 AP를 수행하여 군집 개수 K를 결정한다. 결정된 K값을 이용하여 원데이터에 대해 K-means를 수행하여 초기 군집을 생성한다.

위의 과정으로 만들어진 초기 군집에 대해, 1차로 군집내의 벡터에 대해 계층적 병합 군집화 알고리즘을 적용하여 세부 군집화를 진행한 뒤, 2차로 각각의 세부 군집을 대상으로 다시 계층적 병합 군집화를 진행하여 최종 군집을 생성한다.

2.2 단어 의미 모호성 해소 딥러닝 모델

<그림 2>는 본 연구에서 제안하는 단어 의미 모호성 해소 모델이다. 이는 사전 학습된 언어모델인 BERT[10]를 정밀조정(fine tuning)하는 시퀀스 레이블링(sequence labeling) 모델이다. 입력 문장의 각 단어는 word piece로 분리되어 BERT로 입력되며, BERT의 출력은 transformer encoder layer를 통과시켜 전체적인 문맥 정보를 추가로 얻는다. 이 transformer encoder layer의 출력은 다시 다층 퍼셉트론을 거쳐 softmax를 적용하고 argmax를 취하여 가장 큰 확률값을 갖는 압축된 어휘 의미 태그를 최종적으로 출력한다. 이때 출력 레이블은 분리된 word piece의 첫 번째 토큰을 기준으로 부여한다.



<그림 2> 단어 의미 모호성 해소 딥러닝 모델

3. 실험 및 성능 평가

3.1 실험 환경

본 논문에서는 제안 모델의 성능 평가를 위해 한국어 세종 의미 부착 말뭉치[11]를 사용하였다. 또, 압축 의미 어휘 태그를 추출하기 위해, 한국어 세종 의미 부착 말뭉치의 의미 태그와 같은 의미 번호를 사용하는 표준국어대사전을 사용했다. 특히, 표준국어대사전에 사용된 표제어는 일부에서 원형이나 품사가 없거나 세종 의미 부착 말뭉치와 다르게 표기되어 있어, 이를 전처리로 해결하였다. 즉, 의미 어휘의 품사나 원형은 표준 국어 대사전에 있는 경우, 우선적으로 그것을 사용하고, 없는 경우에 사전의 의미 어휘를 ETRI 형태소 분석기[12]를 이용해 분석한 결과를 추가하여 사용하였다.

말뭉치에 등장한 의미 어휘가 사전과 매칭되지 못하는 경우, 단일 의미 단어로만 구성된 문장, BERT Tokenizer를 거쳤을 때, BERT의 입력길이 512바이트를 초과하는 문장들은 실험에서 사용하지 않았다.

단어 의미 모호성 해소 실험은 시소러스를 이용한 방법과 성능을 비교하기 위해 Vial et al.(2019)[5]의 연구에서 사용한 데이터 크기와 유사한 규모로 39,000문장을 랜덤 샘플링하여 실험에 사용했다. 39,000문장 중 1,060문장은 테스트 데이터로 사용했다. 나머지 37,940

문장을 학습 데이터와 검증 데이터로 나누어 교차 검증을 하였다.

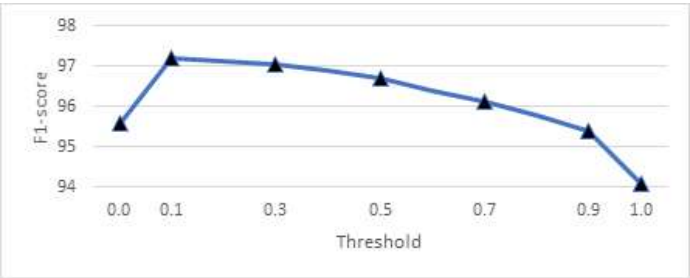
군집화의 하이퍼 파라미터인 유사도 거리 제한 임계값은 0.1부터 1.0까지 다양하게 적용했다. 여기서 유사도 거리는 Euclidean distance로 계산되었기 때문에, 가장 높은 유사도를 1.0으로 전체 유사도 거리를 스케일링하여 군집화에 사용했다.

3.2 실험 결과

모델은 3배수 교차 검증으로 평가되었으며, 성능은 3개 평가 결과의 평균을 계산했다. 이 때, 태그는 의미 어휘 군집화 결과로 생성된 군집 번호로 대체하여 사용하였다. 평가는 [5]의 연구와 마찬가지로 학습에 사용된 의미 태그에 대해서만 F1-score를 계산했다. 딥러닝 모델에 사용되는 BERT는 ETRI에서 제공받은 KorBERT[12]이다. 또한 본 논문에서 진행된 딥러닝 기반 모델의 실험에서 랜덤 시드값은 모두 동일하게 '1314'로 고정하고 진행했다. 모델 실험 결과는 <표 1> 및 <그림 3>과 같다.

<표 1> 다양한 임계값(Threshold)에 따른 성능

Threshold	Sense vocabulary tag size[1)]	Compress. rate	F1-score(%)	Δ
0.0	114,822	0%	95.58	0.0(base)
0.1	104,183	9%	97.21	+1.63
0.3	76,427	33%	97.05	+1.47
0.5	59,864	48%	96.69	+1.11
0.7	39,179	66%	96.12	+0.54
0.9	15,159	87%	95.41	-0.17
1.0	2,410	98%	94.09	-1.49



<그림 3> 군집화 임계값에 따른 모델 성능 변화

군집화하여 압축된 의미 태그를 사용할 경우, baseline 모델인 군집화 이전(임계값 0.0)에 비해 모든 경우에서 성능이 증가함을 볼 수 있다. 특히, 임계값이 0.1일때 F1-score 97.21%로, 압축된 의미 태그 사용 이전에 비해 1.63%p 증가한 가장 높은 성능을 보였으며, 이후의 임계값은 최고점에 비해 성능이 하락한다. 따라서 의미 어휘의 압축률의 높음이 항상 단어 의미 모호성 해소 모델의 높은 성능향상을 보이지 않음을 알 수 있다.

3.3 토론

<표 2>는 [5]의 연구에서 보고한 실험 결과이다. 4개 시스템은 앙상블(ensemble) 기법을 적용한 동일한 딥러닝 모델을 사용하였으며, 압축된 의미 태그 생성방법만 다르다. 여기서 압축 이전에 비해 synonyms 관계를 사용했을 때 성능이 12.62%p 증가했으며, hypernyms 관계를 사용했을 때는 13.33%p, 모든 관계(all relation)에서는 12.00%p 증가하였다. 모든 관계를 사용할 경우, 압축률은 크게 올랐지만, hypernyms 관계만을 사용했을 때에 비해 오히려 그 성능이 떨어졌다. 이는 본 논문의 실험 결과와 같은 경향을 보인다. 즉, <표 1>에서 보듯이 본 논문의 제안 모델에서 임계값이 0.1 보다 커질 경우, 오히려 성능이 점점 떨어지는 경향을 보인다. 따라서 높은 압축률이 반드시 성능 향상을 보이지 않음을 알 수 있으며, 적절한 임계값을 적용하여 의미 어휘를 군집화하여 사용할 필요가 있음을 보인다.

본 논문의 실험결과(<표 1>)와 시소러스를 사용한 [5]의 실험결과(<표2>)를 비교해 보면, 후자의 성능 향상 폭이 높다. 그 이유는 두 가지로 추정된다. 첫째로 [5]의 연구는 수작업으로 구축한 시소러스를 사용하여, 본 연구의 클러스터링 방법보다는 의미 어휘 압축이 더 정확하기 때문으로 추정된다. 둘째로, baseline 모델의 성능차이로 인해 성능 향상 폭이 다르게 보일 수 있다. 즉, 두 성능은 각각 63.40%와 95.58%로, 낮은 성능에서는 상대적으로 더 큰 폭의 성능 향상이 일어날 가능성이 크기 때문이다. 이를 고려하면, 본 연구가 상대적으로 높은 성능인 95.58%에서 1.63%p를 향상 시킨 것은 그 의미가 있다고 판단된다. 또한, 군집화를 이용한 관계 계산의 정확도를 향상시키면 좀 더 성능이 향상될 가능성도 있음을 추측하게 한다.

<표 2> Sense Vocabulary Compression 영어 데이터 성능[5]

System	Vocabulary size	Compres. rate	F1-score avg[2)]	Δ
Baseline	206,941	0%	63.40	0.0
Synonyms	117,659	43%	76.02	+12.62
Hypernyms	39,147	81%	76.73	+13.33
All relation	11,885	94%	75.40	+12.00

4. 결론

단어 의미 모호성 해소 연구는 문장에 등장하는 동형이의어에서 문맥에 맞는 정확한 의미를 결정하는 것으로, 질의응답, 기계번역, 검색 시스템 등 다양한 자연어 처리 응용 시스템에서 보다 정확한 처리를 하기 위한 중요한 과제이다. 이는 지식활용, 지도학습, 비지도학습 등 다양한 방법으로 연구되어 왔으며, 그중 지도학습 방법이 좋은 성능을 보이고 있다. 특히, 지도학습 방법에 지식 정보를 추가로 활용하여, 학습 데이터에서 학습되지 않은 데이터를 효과적으로 처리하는 방법이 많이 연구되고 있다.

최근 연구[5]에서는 WordNet 시소러스를 이용하여 의

1) 단일 의미를 지니는 의미 어휘를 제외한 의미 어휘(동형이의어)의 전체 개수임

2) Senseval, Semeval 5개 데이터 평가 성능의 평균이며, 앙상블 기법을 적용한 모델을 사용했음

미 어휘를 압축하고, 이를 이용하여 단어 의미 모호성 해소하여 우수한 성능을 보였다. 본 연구에서는 시소러스 없이 사전의 단어 의미 정의를 임베딩하고 임베딩된 의미 벡터를 기반으로 다양한 군집화 기법을 활용하여 의미 어휘를 효과적으로 군집화하는 방법을 제안하였다. 군집화된 의미 어휘 태그를 사용하여 세종 의미 부착 말뭉치를 대상으로 실험해본 결과, 사용하지 않을 때보다 F1기준으로 1.63%p 성능이 향상되었다.

향후에는 단어 의미 모호성 해소의 성능을 높이기 위해, 의미 벡터를 효과적으로 생성하는 방법 및 의미 벡터를 다양한 관계로 군집화하는 방법 등을 연구할 예정이다.

감사의 글

이 성과는 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2021R1I1A3059545).

참고문헌

[1] Diana McCarthy. 2009. "Word sense disambiguation: An overview." Language and Linguistics compass 3.2. pages 537-558. 2009.

[2] Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. “GlossBERT: BERT for word sense disambiguation with gloss knowledge.” In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3509-3514. 2019.

[3] Michele Bevilacqua and Roberto Navigli. "Breaking through the 80% glass ceiling: Raising the state of the art in Word Sense Disambiguation by incorporating knowledge graph information." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2854-2864. 2020

[4] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. "SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 05., pages 8758-8765,. 2020.

[5] Loïc Vial, Benjamin Lecouteux, and Didier Schwab. "Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation." arXiv preprint arXiv:1905.05677. 2019.

[6] Global WordNet Association. http://globalwordnet.org.

[7] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope and Ray Kurzweil. 2018. "Universal Sentence Encoder for English." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 169-174, Brussels, Belgium.

[8] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. 2008. "Introduction to Information Retrieval", Cambridge University Press, Cambridge, UK.

[9] Brendan J. Frey and Delbert Dueck. 2007. "Clustering by passing messages between data points." Science, 315(5814):972-976.

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

[11] The National Institute of the Korean Language, “21st Century Sejong Project Final Result, Revised Edition.” 2011.

[12] ETRI Open API Korean Language analysis. http://aiopen.etri.re.kr