단어 의미 모호성 해소를 위한 군집화된 의미 어휘의 품질 향상

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

Improving Clustered Sense Labels for Word Sense Disambiguation

Jeongyeon Park○, Hyeong Jin Shin, Jae Sung Lee
Chungbuk National University

요 약

단어 의미 모호성 해소는 동형이의어의 의미를 문맥에 맞게 결정하는 일이다. 최근 연구에서는 희소 데이터 처리를 위해 시소러스를 사용해 의미 어휘를 압축하고 사용하는 방법이 좋은 성능을 보였다[1]. 본 연구에서는 시소러스 없이 군집화 알고리즘으로 의미 어휘를 압축하는 방법의 성능 향상을 위해 두 가지 방법을 제안한다. 첫째, 의미적으로 유사한 의미 어휘 집합인 범주(category) 정보를 군집화를 위한 초기 군집 생성에 사용한다. 둘째, 다양하고 많은 문맥 정보를 학습해 만들어진 품질 좋은 벡터를 군집화에 사용한다. 영어데이터인 SemCor 데이터를 학습하고 Senseval, Semeval 5개 데이터로 평가한 결과, 제안한 방법의 평균 성능이 기존 연구보다 1.5%p 높은 F1 70.6%를 달성했다.

주제어: 단어 의미 모호성 해소, 군집화, 어휘 압축, 희소데이터

1. 서론

단어의 정확한 의미를 분석하는 단어 의미 모호성 해소 (Word Sense Disambiguation, WSD) 연구는 자연어처리 분야에서 오랫동안 연구되어 왔다. 이 연구는 주로 지도학습, 비지도학습, 지식기반의 3가지 방법이 사용되는데, 특히 지도학습 기반 방법이 가장 높은 성능을 보인다[2,3,4]. 그러나, 지도학습 기반 방법은 학습 데이터 부족 문제로 인한 희소 데이터 처리에 약점을 보인다. 이를 극복하기 위해, 최근 연구에서는 시소러스 등의 지식 정보를 사용한다[1,5,6,7,8]. 여기에는 의미 어휘를 압축하고 사용하는 **의미 어휘 압축 방법**[1,5], 문맥을 직간접적으로 비교하는 **문맥 비교 방법**[6,7,8] 등이 있다.

의미 어휘 압축 방법 중, [1] 연구에서는 시소러스에 나타난 단어의 상하위어 관계를 이용하여 의미 어휘(sense vocabulary)를 압축(Sense Vocabulary Compression, SVC)하고, 이를 단어 의미 모호성 해소 모델에 사용했다. [5] 연구는 의미 어휘의 사전 정의(sense definition)를 벡터로 인코딩하고 군집화(clustering) 알고리즘을 사용해 의미 어휘를 압축했다.

문맥 비교 방법 중, [6]은 동형이의어가 포함된 문장을 사전의 의미 정의와 비교하는 GlossBERT 모델을 제안했다. [7]은 문맥과 사전정의를 따로 인코딩하여 유사도를 계산하는 BEM(Bi-Encoder Model) 모델을 제안했다. [8]은 동형이의어의 의미 정의 후보 전체를 함께 인코딩하고, 정답이 될 수 있는 의미 정의의 범위를 추출하는 방법으로 의미를 결정하는 ESCHER(transformer based Extractive Sense Comprehension) 모델을 제안했다.

의미 어휘 압축 방법은 문맥 비교 방법에 비해 비교적 낮은 성능을 보이지만, 파라미터 수를 줄여 빠른 처리 속도로 희소 데이터를 효과적으로 처리할 수 있다. 왜냐하면, 문맥 비교 방법은 문장 하나에 포함된 n개의 동형이의어 처리를 위해 n번의 인코딩 과정이 필요하기 때문이다. 반면, 의미 어휘 압축 방법은 시퀀스 레이블링(sequence labeling) 모델을 사용하므로, 하나의 문장에 포함된 n개의 동형이의어 처리를 위해 한 번의 인코딩 과정만 거치면 된다.

본 연구에서는 [5]에서 사용한 연구방법(이하, **의미 어휘 군집화**)의 성능 향상을 위해 다음의 방법을 사용한다. 첫째, 유사한 의미를 가진 단어를 모아놓은 범주 정보인 워드넷(WordNet)[9]의 Super Sense 집합(이하, SS)을 군집화의 초기 군집 생성에 사용한다. 둘째, 다양하고 많은 문맥을 수집해 만든 품질 좋은 벡터인 ARES(context-AwaRe Embeddings for Senses)[10] 벡터를 사용한다.

2. 의미 어휘 군집화 모델

2.1 의미 어휘 군집화

사전 정의를 이용한 의미 어휘 군집화(Sense Definition Clustering, **SDC**)의 전체 과정은 다음과 같다[5]. 먼저, 사전의 의미 정의를 Universal Sentence Encoder[11]를 사용해 벡터로 표현(Sense Definition Vector, **SDV**)한다. 그런 뒤, HAC(Hirachical Agglomerative Clustering)[1)][12]에서 군집 내에 동형이의어가 발생하지 않도록 하는 조건을 추가하여 군집화한다. 이 때, 유클리드 거리(Euclidean distance)로 계산된 유사도 거리 중, 가장 높은 유사도 거리값을 1.0으로 전체 유사도 거리를 조정하고 군집화한다. 그리고 HAC의 $O(n^3)$ 시간복잡도로 인한 처리 속도를 완화하기 위해, 보조 군집화 알고리즘을 사용해 어느 정도 군집화된 초기 군집을 만들어 사용한다.

[5]의 연구에서, SDC를 위한 **초기 군집 생성 과정**은

1) Group average agglomerative clustering 방법을 사용했음

다음과 같다. 먼저, 의미 어휘를 **품사 기준**으로 분류한다. 그리고, 분류된 각각의 군집을 유사도 전파(Affinity Propagation, 이하 AP)[2][13] 알고리즘을 사용해 군집화한다. 만약, AP를 실행할 수 없는 환경이라면, AP를 실행할 수 있는 최대한의 데이터를 랜덤으로 추출한 뒤, AP를 사용해 군집 개수를 정한다. 그런 다음, 결정된 군집 개수와 K-means[3][12]를 사용해 군집화한다.

위 과정으로 만들어진 각각의 군집에 SDC를 사용해 군집 내에 동형이의어가 없는 초기 군집들을 생성한다. 이렇게 만들어진 초기 군집 전체를 대상으로 SDC를 사용해 군집화된 레이블을 생성한 뒤, 단어 의미 모호성 해소에 사용한다.

2.2 단어 의미 모호성 해소 모델

의미 어휘 군집화로 생성된 레이블은 단어 의미 모호성 해소를 위해 사용된다. 단어 의미 모호성 해소 모델은 [그림 1]과 같은 BERT[14] 기반 딥러닝 모델이다.

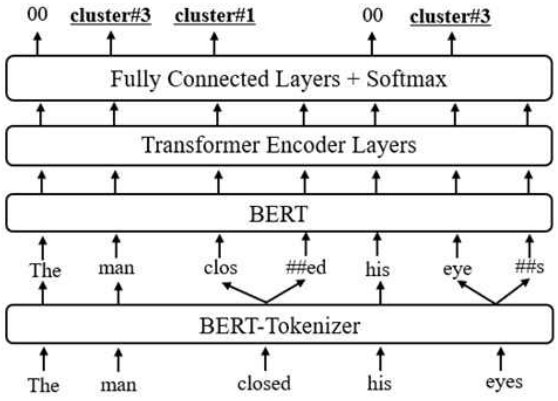

그림 1. BERT 기반의 WSD 모델[5]

3. 제안 방법

[15,16] 연구 결과에 따르면, 사용되는 벡터가 군집화 결과에 큰 영향을 미친다. 특히, 데이터를 잘 표현하는 벡터를 생성하고, 생성된 벡터를 군집화에 사용했을 때 최종 평가 결과가 좋았다. 따라서, 의미 어휘를 잘 표현하는 벡터를 SDC에 사용하면, 단어 의미 모호성 해소 모델의 성능도 향상될 수 있다.

또한, SDC에서는 보조 군집화 알고리즘을 사용해 어느 정도 군집화된 초기 군집을 만들어 사용했다. 그렇다면, 초기 군집 생성 결과가 최종적으로 군집화된 레이블 생성에 영향을 줄 것이다. 따라서, 본 연구에서는 의미 어휘 군집화 모델[5]에서 군집화된 레이블의 품질을 높이기 위해, 두 가지 방법(SS+SDV, SS+ARES)을 사용하는 모델을 제안한다.

1) SS+SDV 방법: 이 방법은 SDC에 사용할 초기 군집 생성을 위해, 전체 데이터를 품사 기준으로 분류하는 대신, 의미적으로 유사한 단어들을 수작업으로 분류한 레이블을 사용한다. 이러한 레이블은 사전 내에 범주 분류(예를 들어, 정보통신, 법률, 건축 등의 분야)로 되어있는 경우가 많다. 대표적으로, 한국어 사전인 우리말샘[4]이나, 표준국어대사전[5]에도 단어마다 의미를 고려해 범주 분류가 되어 있으며, 영어 사전인 워드넷에도 **Super Sense(SS)** 라는 이름으로 범주 분류가 되어 있다. 본 연구에서는 SDC의 초기 군집 생성을 위해 SS를 사용한다. SS의 일부 예시는 아래 [표 1]과 같다.

표 1. Super Sense 집합(SS) 예시

집합 이름	단어 예시
noun.animal	dog, fish, cow, horse, human
noun.person	player, actor, author, client, parent
verb.emotion	love, appeal, charm, hate, trust
verb.motion	move, crash, bring, swim, escape

2) SS+ARES 방법: 이 방법은, 초기 군집 생성에 SS를 사용하면서, 사전의 의미 정의만 사용해 생성된 벡터인 SDV 대신, 더 많은 정보로 학습된 **ARES 벡터**를 군집화에 사용한다. ARES 벡터는 의미 어휘의 사전 정의 외에, 여러 문서에서 해당 의미 어휘가 사용된 문맥 정보를 수집해 만들어졌다[10]. 또한, 사전에 등재된 의미 정의 외에, 유의어 관계에 있는 단어들이 사용된 문맥을 여러 개 추출해서 벡터 생성에 사용했다.

4. 실험 및 결과

4.1 실험 환경

실험은 영어 데이터를 사용하여 진행한다. 실험에 사용된 데이터 중, 모델의 학습에는 SemCor 3.0을 사용했으며, 평가 데이터는 Senseval, Semeval의 5개 데이터를 사용했다. 군집화는 유사도 거리 제한을 0.1부터 1.0까지 다양하게 적용해 진행했으며, 군집화 과정에서 SS와 ARES 벡터를 각각 사용했다. WSD 모델은 의미 어휘 압축 결과를 사용하는 [그림 1]과 같은 딥러닝 모델이다. 딥러닝 모델에 사용된 하이퍼 파라미터는 [표 2]와 같다.

표 2. 딥러닝 모델 하이퍼 파라미터

learning rate	2e-5
embed size	768
dropout rate	0.1
transformer layer num	6
transformer layer dim	2048
attention head num	8

모델의 비교를 위해, 의미 어휘 압축 방법을 사용하지 않는 **기준 모델**(baseline), [5]의 의미 어휘 군집화 방법을 사용한 **POS+SDV** 모델, 본 연구의 방법을 사용한 **SS+SDV**, **SS+ARES** 모델을 실험했다.

2) 적절한 군집 개수를 자동으로 결정하는 군집화 알고리즘
3) 주어진 K 개의 군집으로 군집화하는 알고리즘
4) https://opendict.korean.go.kr/
5) https://stdict.korean.go.kr/

4.2 실험 결과

평가는 모델을 사용해 동형이의어의 의미 레이블을 결정하고 F1-score를 계산했다. 여기서, 실험 결과의 성능은 5개 데이터를 각각 평가하고 평균을 계산한 것이다. 이 때, 기준 모델을 제외하고, 본래의 주석된 의미 레이블을 군집화된 레이블로 대체하여 사용했다. 실험 결과는 [표 3]과 같다.

표 3. 모델에 따른 최대 성능 비교

의미 어휘 압축 방법	모델	의미 어휘 태그 개수	의미 어휘 압축률	F1-score (%)	성능 변화 (%p)
압축 없음	기준 모델 (baseline)	207K	0%	62.9	0.0
군집화	POS+SDV [5]	22K	89%	69.1	+6.2
	SS+SDV (제안모델)	85K	59%	69.4	+6.5
	SS+ARES (제안모델)	49K	77%	70.6	+7.7
시소러스	SVC[6][1]	39K	81%	75.7	+12.8

기준 모델은 F1-score 기준 62.9%의 성능을 보이고, [5] 연구결과인 **POS+SDV** 모델의 성능은 기준 모델보다 F1-score 기준 6.2%p 높은 69.1% 성능을 보였다. 그리고 **SS+SDV** 모델은 69.4%, **SS+ARES** 모델은 70.6%의 성능을 각각 달성했다. 이것은 의미적으로 잘 분류된 범주 정보를 군집화에서 초기 군집으로 사용하거나, 데이터를 잘 나타내는 벡터를 사용하는 것이 높은 품질의 군집화된 레이블을 생성할 수 있음을 보인다.

본 논문의 제안 방법은 잘 구축된 시소러스의 단어간 관계를 사용한 SVC[4] 모델에 비해 낮은 성능을 보인다. 하지만, 군집화된 레이블의 품질을 높여 WSD 모델의 성능 향상을 보였다. 따라서 군집화에 사용되는 자원에 따라 추가적인 성능 향상을 기대할 수 있다.

5. 결론

단어 의미 모호성 해소 연구는 단어의 의미를 문맥에 맞도록 결정하는 중요한 과제다. 이는 질의응답, 기계번역 등 다양한 자연어처리 응용 시스템의 성능 향상에 기여할 수 있다.

최근 단어 의미 모호성 해소 연구는 기존의 지도 학습 기반 방법의 한계를 극복하고자 시소러스 등의 지식 정보를 추가로 사용하는 방법을 사용하고 있다. 이 중, 의미 어휘 압축 방법을 사용한 [5] 연구에서는, 의미 어휘를 압축하기 위해 군집화 기법을 사용했다.

본 연구에서는 군집화된 의미 어휘의 품질 향상을 위해 아래와 같이 두 가지 방법을 사용했다. 1) 군집화의 초기 군집으로 유사한 의미를 지닌 단어를 모아놓은 Super Sense 집합을 사용하고, 2) 품질 좋은 벡터인 ARES 벡터[13]를 군집화에 사용했다. 그런 뒤, 단어 의미 모호성 해소 모델의 성능 변화를 확인하기 위해, 영어 데이터인 SemCor 말뭉치를 학습하고, Senseval, Semeval 5개 데이터를 각각 평가한 뒤, 평균을 계산하여 이전 모델과 비교했다. 평가 결과, 본 논문에서 제안한 방법이 의미 어휘 압축 방법을 사용하기 전인 기준 모델 성능 F1 62.9% 보다 7.7%p 높고, 기존의 의미 어휘 군집화 방법[5] 보다 1.5%p 높은 F1 70.6% 성능을 보였다.

6) 어휘간 상하위어 관계를 이용해 압축된 레이블을 사용한 모델

감사의 글

이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임(No. 2021R1I1A3059545).

참고문헌

[1] L. Vial, L. Benjamin, and S. Didier, "Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation," *arXiv preprint arXiv:1905.05677*. 2019.

[2] R. Navigli, "Word sense disambiguation: A survey," *ACM computing surveys (CSUR) 41(2)*, pp.1-69, 2009.

[3] D. McCarthy, "Word sense disambiguation: An overview,", *Language and Linguistics compass 3(2)*, pp.537-558, 2009.

[4] J. Sreedhar, S. V. Raju, A. V. Babu, A. Shaik, and P. P. Kumar, "Word sense disambiguation: An empirical survey," *International Journal of Soft Computing and Engineering (IJSCE) ISSN*, pp. 2231-2307, 2012.

[5] J. Y. Park, H. J. Shin, and J. S. Lee. "Word Sense Disambiguation Using Clustered Sense Labels," Applied Sciences 12(4): 1857, 2022.

[6] L. Huang, C. Sun, X. Qiu, and X. 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*, pp. 3509-3514, 2019.

[7] T. Blevins, and L. Zettlemoyer, "Moving down the long tail of word sense disambiguation with gloss-informed biencoders," arXiv preprint arXiv:2005.02590, 2020.

[8] E. Barba, T. Pasini, and R. Navigli, "ESC: Redesigning WSD with extractive sense comprehension," Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021.

[9] C. Fellbaum, "WordNet," Theory and applications of ontology: computer applications. Springer, Dordrecht, pp.231-243, 2010.

[10] B. Scarlini, T. Pasini, and R. Navigli, "With more contexts comes better performance: Contextualized sense embeddings for all-round word sense disambiguation," Proceedings of the

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.3528-3539, 2020.
[11] D. Cer, Y. Yang, S. Kong, S. N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope and R. Kurzweil, "Universal Sentence Encoder," In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 169-174, 2018.
[12] C. D. Manning, P. Raghavan and H. Schütze, "Introduction to Information Retrieval," Cambridge University Press, Cambridge, UK, 2008.
[13] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," Science, 315(5814), pp.972-976, 2007.
[14] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[15] M. Dash and H. Liu, "Feature selection for clustering," Pacific-Asia Conference on knowledge discovery and data mining, Springer, Berlin, Heidelberg, pp.110-121, 2000.
[16] J. Miao and L. Niu. "A survey on feature selection," Procedia Computer Science 91, pp.919-926, 2016.