워드 임베딩을 이용한 세종 전자사전 확장

박다솔○, 차정원
창원대학교
dasol_p@changwon.ac.kr, jcha@changwon.ac.kr

Extension Sejong Electronic Dictionary Using Word Embedding

Da-Sol Park○, Jeong-Won Cha
Changwon National University

요 약

본 논문에서는 워드 임베딩과 유의어를 이용하여 세종 전자사전을 확장하는 방법을 제시한다. 세종 전자사전에 나타나지 않은 단어에 대해 의미 범주 할당의 시스템 성능은 32.19%이고, 확장한 의미 범주 할당의 시스템 성능은 51.14%의 성능을 보였다. 의미 범주가 할당되지 않은 새로운 단어에 대해서도 논문에서 제안한 방법으로 의미 범주를 할당하여 세종 전자사전의 의미 범주 단어 확장에 대해 도움이 됨을 증명하였다.

주제어: 워드 임베딩, 세종 전자사전, 의미 범주

1. 서론

의미 분석은 문장을 구성하는 단어들의 의미를 구분하고, 문장 구성 성분들 사이의 의미적 관계를 논리적으로 밝혀내어 문장의 전체적 의미를 파악하는 기술을 말한다.

의미 분석은 형태소 분석과 구문 분석의 과정을 거쳐 이루어지는 자연어 처리의 상위 단계이다[1]. 크게 두 가지로 중의성을 해소하는 문제와 의미역을 결정하는 문제로 나눌 수 있다.

의미역 결정에는 의미 논항 역할 정보(Semantic Role)와 의미 범주 정보를 사용해야 한다. 하지만 의미 논항 역할 정보와 의미 범주 정보가 포함되어 있는 세종 전자사전의 데이터는 한국어 문장을 처리하기 위해 확장할 필요가 있다. 본 논문에서는 세종 전자사전의 확장을 시도한다.

본 논문에서 1.2장에서는 세종 전자사전에 대해 소개한다. 2장에서는 관련연구에 대해 소개한다. 제안한 방법에 대해서는 3장에서 기술하고, 마지막으로 4장에서는 결론을 기술한다.

1.2. 세종 전자사전

본 연구에서 사용된 21세기 세종 계획의 전자사전은 현대 한국어 어휘 전반에 대한 종합적이고 방대한 정보를 담고 한국어 자동 처리에 보편적으로 사용될 수 있으며 다양한 전산 처리에 필수적이고 실용적인 전자사전이다[2].

세종 전자사전은 표제어에 대해 다양한 통사적, 의미적 정보가 XML형태로 수록되어 있으며 본 논문에서 의미역을 결정하는데 사용한 격틀 정보가 포함되어 있다. 그 중 격틀 정보를 추출하여 사용하였다. 세종 전자사전은 25,458개의 명사, 15,181개의 동사, 4,398개의 형용사와 645개의 명사 의미 하위 부류와 631개의 용언 의미 하위 부류로 구성되어 있다.

2. 관련 연구

코퍼스로부터 상·하위어를 추출하는 연구들이 있었다. Hearst는 텍스트에서 패턴 인식과 의미관계를 이용하여 상·하위어 의미 관계를 자동적으로 추출하는 방법을 처음으로 제안하였다[3]. 여러 가지 문법적 패턴을 생성한 후, 주어진 문장의 형태가 패턴 형태와 동일하면 관계 트리플을 추출한다. 동일한 패턴이지만 수식어들이 포함되어 있는 경우에는 수식어를 걸러내지 못한다는 문제점이 있다.

Cederberg, Widdows는 텍스트에서 상·하위어 관계를 자동 추출하는 데 있어 “Latent Semantic Analysis(LSA)”를 사용하였다[4]. LSA를 사용함으로써 정확률과 재현율을 향상 시켰다.

Verginica은 상·하위어가 공기(co-occurrence)하는 패턴을 확인하고 상·하위어 관계의 패턴 생성에 대한 연구를 하였다[5]. 그러나 부사어나 관사가 포함된 문장은 공기하는 패턴에 적용되지 않는다는 문제점이 있다.

Erik Tjong Kim Sang, Katja Hofmann and Maarten de Rijke는 코퍼스의 어휘 패턴과 의존 패턴을 이용하여 상·하위어를 추출하였다[6].

Marco Baroni, Bgoc-Quynh Do and Chung-chieh Shan은 구의 분포적 벡터 표현을 이용하여 형용사-명사 구조와 한정사-명사 구조의 함의를 찾는 연구를 진행하였다[7]. 형용사-명사 구조와 한정사 또한 의미적 벡터로 표현되어 있고 분포적 벡터로 SVM, classifier를 이용하여 함의를 찾을 수 있었다.

Marek Rei, Ted Briscoe는 벡터로 하위어를 찾는 연구를 진행하였다[8]. 패턴 기반의 하위어를 찾는 것은 함께 언급되는 두 단어에만 의존하기 때문에 매우 낮은 재현율을 가져온다. 지도학습 또는 패턴 구조 없이 다른 도메인과 다른 언어에도 적용할 수 있는 벡터 유사도 방법을 이용하였다. 의존성 기반 벡터 표현을 사용하여 신경 네트워크와 윈도우 기반의 모델을 사용하여 최고 성능을

달성했다.

한국어의 상·하위어 추출 연구에는 방찬성, 이해윤은 코퍼스를 이용하여 상·하위어 관계 패턴을 추출하는 방법을 제안하였다[9]. 패턴 추출을 목적으로 명사의 열거를 나타낼 때 다양한 조사 또는 문장부호의 변이로 인하여 고정된 패턴 포착에 대한 어려움이 있다. 그리고 문맥에 의존적인 어휘들이 나타날 때 어휘 자체만으로 상·하위어 판단이 어렵다. 최유미, 사공철은 자동 시소러스 구축을 위한 상위어 자동 추출을 연구했다[10]. 문헌정보학 용어사전에 기술된 문장의 구문적 특성을 조사하였다. 그리고 표본조사를 통하여 얻은 구문정보를 이용하여 10개의 알고리즘을 개발하였으며, 89.4%의 정확도를 보였다.

3. 제안방법

본 논문에서는 세종 전자사전에 나타나지 않는 새로운 단어의 의미 범주를 할당하기 위해서 할당하고자 하는 단어의 임베딩 벡터와 유의어를 사용한다. 먼저 대용량 문서에서 기존 세종 전자사전에 있는 단어들의 임베딩을 구했다. 워드 임베딩은 구글의 word2vec[11]을 사용하였다. 이 값과 세종 전자사전에 나타나지 않는 단어들의 유사도를 구하여 새로운 단어의 어휘 범주를 할당한다. 또한 세종 전자사전에 있는 단어의 유의어와 세종 전자사전에 나타나지 않는 유의어를 비교하여 새로운 단어에 의미 범주를 할당한다. 모든 실험의 cosine similarity와 pearson 상관관계는 높은 값을 가지는 단어의 의미 범주를 할당하고 euclidean distance는 낮은 값을 가지는 단어의 의미 범주를 할당한다. 선형 조합을 적용할 때 가중치 값들은 0.1 단위로 조절하며 모든 경우를 실험하여 결정하였다.

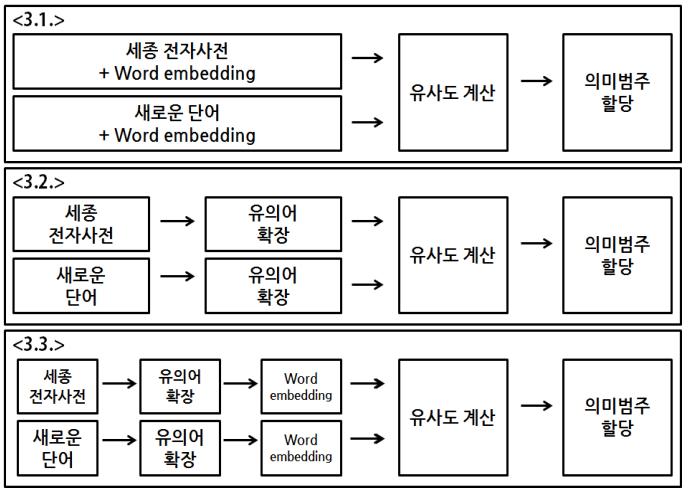

그림 1. 제안 방법

그림 1은 본 논문에서 제안한 방법을 요약하여 보여준다. 이 세 방법에 대해서는 3.1, 3.2, 3.3장에서 자세하게 설명한다.

3.1. 어휘 유사도를 이용

첫 번째로 시도한 방법은 워드 임베딩 벡터를 이용하여 의미 범주를 할당하는 방법이다. 대용량 문서를 이용하여 세종 전자사전의 단어들에 대한 워드 임베딩 값을 구한다. 또한 의미 범주를 할당하고자 하는 단어들도 같은 방법으로 워드 임베딩 값을 구한다. 이 단어들 사이의 임베딩 값의 유사도를 계산하여 의미 범주를 할당한다.

단어들의 임베딩 값을 구하기 위해서 인터넷 신문과 한국어 위키피디아에서 모은 2억 8천만 형태소 코퍼스를 사용하였다. 이 코퍼스를 창원대학교 형태소 품사 태거[12]를 사용하여 품사를 부착하였다.

유사도 계산을 위해서 cosine similarity[13,14,15], euclidean distance[16], pearson 상관관계[17]를 사용하였다. best 1은 유사도 가장 높은 하나가 정답일 때이고 best 5는 유사도 높은 상위 5개 중에 정답이 있을 때의 정확도이다. 표 1은 실험 결과를 보여준다. 테스트에는 새로운 단어 876개를 사용하였다.

표 1. 실험 결과

유사도	best 1	best 3	best 5
cosine similarity	25.57% (224/876)	42.57% (373/876)	49.77% (436/876)
euclidean distance	25.00% (219/876)	42.57% (373/876)	49.77% (436/876)
pearson 상관관계	25.45% (223/876)	42.35% (371/876)	50.68% (444/876)

3.1.1. 실험에 대한 고찰

단어의 의미 유사도를 구하기 위해서 워드 임베딩 값을 사용하였다. 표 2는 범주를 할당할 단어들을 보여준다. 표 2를 보면 워드 임베딩으로 계산된 유사도는 상위 유사도와 하위 유사도를 구분하지 못하는 경우가 있다는 것을 볼 수 있다. 이것은 워드 임베딩이 자세한 의미 구분에는 한계가 있다는 것을 보여준다. 또한 저빈도 단어에 대해서는 정확한 값을 구하지 못하는 경우도 발견할 수 있었다.

표 2. 범주를 할당할 단어

세종 전자사전 단어	바톤/NNG	삐에로/NNG	각각/NNG
best 5 단어	콜/NNG 큐/NNG 하프/NNG 헤드/NNG 러브/NNG	고백/NNG 제비족/NNG 짝사랑/NNG 진짜/NNG 무심/NNG	형태/NNG 방식/NNG 각자/NNG 순서/NNG 특징/NNG

3.2. 유의어 확장을 통한 매칭 이용

단어 자체의 임베딩 값의 단점을 보완하기 위해서 단어의 유의어를 이용하여 유사도를 계산한다. 여기에 사용한 유의어는 대국어사전의 유의어 정보, 개체명 사전

정보, 워드 임베딩을 이용한 의미 유사어 정보 등이다. 확장한 단어들의 매칭 수를 두 번째 자질로 사용한다. 어휘 유사도(α)와 유의어를 확장한 단어의 매칭 수(β)를 선형 조합한다. 표 3은 실험 결과를 보여준다. 표 4는 선형 조합에 적용된 가중치 값이다.

표 3. 실험 결과

유사도	best 1	best 3	best 5
cosine similarity	31.96% (280/876)	49.09% (430/876)	56.96% (499/876)
euclidean distance	4.57% (40/876)	12.67% (111/876)	18.49% (162/876)
pearson 상관관계	31.62% (277/876)	48.97% (429/876)	56.96% (499/876)

표 4. 선형조합 가중치 값

유사도	α	β
cosine similarity	0.7	0.3
euclidean distance	0.1	0.9
pearson 상관관계	0.7	0.3

3.2.1. 실험에 대한 고찰

유의어를 사용할 경우 유사도를 사용할 때 워드 임베딩 벡터 값이 존재하지 않는 단어에 대해서 평가를 할 수 없다. 또한 유사도 측정에서 상대적으로 낮은 값이었던 정답이 유의어 사전을 통해서 더 높은 값을 가지게 되어 순위 변동이 있고, 정답으로 선택되었음을 알 수 있다. 하지만 유의어의 어휘 단순 매칭만으로는 의미 범주 할당에 대한 뒷받침이 부족하다.

3.3 유의어 확장 및 워드 임베딩 이용

유의어로 얻을 수 있는 추가 정보인 유의어의 유사도를 계산한다. 확장한 단어들의 유사도를 세 번째 자질로 사용한다. 어휘 유사도(α)와 유의어 매칭 수(β) 그리고 어휘 유사도와 유의어를 확장한 단어의 유사도(γ)를 선형 조합한다. 표 5는 실험 결과를 보여준다. 표 6은 선형 조합에 적용된 가중치 값이다.

표 5. 실험 결과

유사도	best 1	best 3	best 5
cosine similarity	32.19% (282/876)	49.43% (433/876)	57.76% (506/876)
euclidean distance	16.44% (144/876)	28.88% (253/876)	36.42% (319/876)
pearson 상관관계	31.96% (280/876)	49.32% (432/876)	57.88% (507/876)

표 6. 선형조합 가중치 값

유사도	α	β	γ
cosine similarity	0.4	0.3	0.3
euclidean distance	0.1	0.4	0.5
pearson 상관관계	0.4	0.3	0.3

3.3.1. 실험에 대한 고찰

유의어 확장과 워드 임베딩을 이용한 실험 성능 향상을 확인할 수 있었다. 이는 단어의 유사성을 나타내는 벡터보다 의미를 나타내는 자질인 유의어에 의한 것이다.

표 7은 단어와 단어에 대한 의미 범주 예시이다. 표 7의 세종 전자사전 단어와 새로운 단어가 유사한 의미 범주나 동일한 의미 범주를 가져야한다고 예상된다. 그러나 실제 의미 범주는 동일하지 않다. 현재 실험은 의미 범주의 어휘가 동일할 때에만 정답으로 처리한다. 하지만 현재 단어의 의미 범주만 정답으로 간주하지 않고, 해당 의미 범주의 부모 의미 범주까지 정답으로 간주해야 한다고 판단하였다.

표 7. 단어와 의미 범주의 예

세종 전자사전 단어		새로운 단어	
어휘	의미 범주	어휘	의미 범주
의료기관/NNG	기관	금융기관/NNG	금융기관
재혼/NNG	만남	결혼/NNG	대칭적행위
개발도상국/NNG	상황값	선진국/NNG	국가

3.4 부모 의미 범주 이용

위의 분석 결과를 바탕으로 추가 실험을 진행하였다. 3.3 실험의 결과에 부모 의미 범주까지 적용하였다. 표 8은 실험 결과이고 표 9는 선형 조합을 할 때 적용한 가중치 값이다.

표 8. 실험 결과

유사도	best 1	best 3	best 5
cosine similarity	51.14% (448/876)	71.58% (627/876)	78.20% (685/876)
euclidean distance	30.37% (266/876)	49.66% (435/876)	57.88% (507/876)
pearson 상관관계	51.26% (449/876)	70.55% (618/876)	78.31% (686/876)

표 9. 선형조합 가중치 값

유사도	α	β	γ
cosine similarity	0.2	0.5	0.3
euclidean distance	0.1	0.4	0.5
pearson 상관관계	0.1	0.5	0.4

4. 결론

본 논문에서는 워드 임베딩과 유의어를 사용하여 세종 전자사전을 확장하는 방법을 제안하였다. 세종 전자사전에 있는 단어와 새로운 단어의 워드 임베딩 유사도를 이용하여 새로운 단어의 어휘 범주를 할당한다. 또한 세종 전자사전에 없는 단어의 유의어와 새로운 단어의 유의어를 비교하여 의미 범주를 할당한다. 부모 의미 범주를 이용한 실험의 best 5 결과에 정답인 단어를 해당 의미 범주에 추가하는 과정을 거쳐 의미 범주를 확장하였다.

실험 결과에 의하면, 어휘 유사도와 유의어 확장 및 워드 임베딩 자질이 세종 전자사전의 의미 범주를 할당하기 위한 단어를 찾아내는 데 도움이 되었다. 그러나 어휘 유사도 자질은 단어 유사도를 찾는 데에는 도움이 되지만 어휘 의미를 구별하기에는 어려움이 있다. 유의어 자질은 의미 범주를 할당하는 데 도움이 되었다.

워드 임베딩을 이용하여 벡터로 표현된 형태소는 다양한 의미들이 포함된 벡터를 생성한다. 그 벡터의 의미 범주를 할당하여도 다양한 의미 중 어떤 뜻을 나타내는 단어인지 알 수 없다는 문제점이 있다. 따라서 다양한 의미를 가지는 단어의 벡터에서 중의성을 가려낼 방법이 필요하다. 그 문제점을 해결하기 위해 단어의 중의성을 해결할 수 있는 단어 벡터 클러스터링 방법에 대한 연구를 진행할 계획이다.

참고문헌

[1] Daniel Jurafsky, James H. Martin, “Speech and Language Processing”, vol.2, pp.546, 2007.

[2] 국립 국어원, “21세기 세종계획 최종 성과물”, 문화체육관광부, 2010.

[3] Hearst,M.A., “Automatic Acquisition of Hyponyms from LargeText Corpora”, Association for Computational Linguistics, pp.539-545, 1992.

[4] Cederberg,S. and Widdows,D., “Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponymy Extraction”, Proc.of the Conference on Natural Language Learning-2003, pp.111-118, 2003.

[5] Verginica Barbu Mititelu, “Automatic Extraction of Patterns Displaying Hyponym-Hypernym Co-Occurrence from Corpora”, Proceedings of First Central European Student Conference in Linguistics, 2003.

[6] SANG, Erik Tjong Kim; HOFMANN, Katja; DE RIJKE, Maarten. “Extraction of Hypernymy Information from Text”, Interactive Multi-modal Question-Answering, pp. 223-245, 2011.

[7] Baroni, Marco, et al. "Entailment above the word level in distributional semantics.", Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp.23-32, 2012.

[8] Rei, Marek, and Ted Briscoe.,"Looking for Hyponyms in Vector Space“,the Conference on Natural Language Learning ,pp.68-77, 2014.

[9] 방찬성,이해윤, “코퍼스를 이용한 상하위어 추출 연구”,인지과학 19.2, 2007.

[10] 최유미,사공철, "상위어 자동추출 알고리즘 개발." 한국정보관리학회 제 5 회 학술대회 논문집 , pp.227-230,1998.

[11] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space.", CoRR, abs/1301.3781, 2013.

[12] URL: https://air.changwon.ac.kr/~airdemo/kg_tagger/ ,2016-09-10.

[13] Gerard Salton, A. Wong, and C. S. Yang. A vector space model for information retrieval. Communications of the ACM, pp.613-620, 1975.

[14] Singhal, Amit. "Modern information retrieval: A brief overview.", IEEE, pp.35-43. 2001.

[15] Manwar, A. B., et al. "A Vector space model for information retrieval: A MATLAB approach." Indian Journal of Computer Science and Engineering (IJCSE), pp.222-229, 2012.

[16] Danielsson, Per-Erik. "Euclidean distance mapping.", Computer Graphics and image processing, pp.227-248, 1980.

[17] Pearson K., “Notes on the history of correlation”, Biometrika, vol.13, pp.25-45, 1920.