Word2Vec를 이용한 단어 의미 모호성 해소

강명윤○, 김보겸Ψ, 이재성†
충북대학교, 비즈니스데이터융합학과○, 디지털정보융합학과Ψ, 소프트웨어학과†
myungyun@chungbuk.ac.kr○, bogyum@chungbuk.ac.krΨ , jasonlee@chungbuk.ac.kr†

Word Sense Disambiguation using Word2Vec

Myung Yun Kang○, Bogyum KimΨ, Jae Sung Lee†
Dept. of business data convergence○ , Dept. of digital informatics and convergenceΨ, Dept. of computer sciences, Chungbuk national university†

요 약

자연어 문서에 출현하는 단어에는 중의적 단어가 있으며, 이 단어에서 발생되는 의미 모호성은 대개 그 문맥에 따라 해소된다. 의미 모호성 해소 연구 중, 한국어 단어 공간 모델 방법은 의미 태그 부착 말뭉치를 이용하여 단어의 문맥 정보를 구축하고 이를 이용하여 모호성을 해결하는 연구로서 비교적 좋은 성능을 보였다. 본 연구에서는 Word2Vec를 이용하여 기존 연구인 한국어 단어 공간 모델의 단어 벡터를 효과적으로 축소할 수 있는 방법을 제안한다. 세종 형태 의미 분석 말뭉치로 실험한 결과, 제안한 방법이 기존 성능인 93.99%와 유사한 93.32%의 정확률을 보이면서도 약 7.6배의 속도 향상이 있었다.

주제어: 단어 의미 중의성 해소, Word2Vec, 동형이의어, 단어 공간 모델

1. 서론

자연어 문서에 출현하는 단어는 중의성을 가질 수 있으며, 단어의 정확한 의미 파악은 정보검색이나 기계번역 등의 자연어 처리 시스템의 성능을 높일 수 있다. 예를 들어 어절 '배다' 에서 '배' 는 품사가 명사, 동사 등이 될 수 있으며, 품사가 명사일 경우에는 먹는 배, 운송수단 배, 신체부위 배 등으로 여러 가지 의미가 있으며, 동사일 경우는 냄새가 스며드는 '배다', 배속에 새끼를 가지는 '배다' 등의 의미를 가진다. 따라서 이 단어의 뜻에 맞는 문서를 찾아주거나 번역을 하기 위해서는 단어의 의미 구분이 선행되어야 한다. 일반적으로 단어 중의성 해소는 품사 태깅이 된 단어에 대해 의미 모호성을 해소하는 것이며, 본 논문에서도 이런 가정을 따른다.

한국어 단어 의미 중의성 해소는 다양한 방법으로 많은 연구가 진행되어 왔다[1-6]. 이 중 단어 공간 모델에 기반을 둔 방법은 대상 단어의 문맥에 나온 또 다른 단어들을 벡터의 각 요소로 표현하여 대상 단어들의 의미를 비교하는 방법이다. [1]에서는 세종 형태 의미 분석 말뭉치를 학습하여 중의성 있는 단어들에 대한 한국어 단어 공간 모델[7]을 만들고, 이를 이용하여 의미 모호성을 해소하여 우수한 성능을 보였다. 하지만 이 방법은 의미 단어 벡터의 차원이 너무 커져 공간이 많이 필요하고 처리 속도가 느려지는 단점이 있다. 이러한 단점을 보완하기 위해 본 연구에서는 Word2Vec[8]을 활용하여 성능을 유지하면서도 단어 벡터의 차원을 효과적으로 줄이는 방법을 제안한다.

2. 관련 연구

기존 단어 의미 중의성 해소 연구는 여러 가지 방법으로 연구가 진행되었다. 이러한 방법에는 사전에 나타난 뜻풀이 글 등을 이용하는 사전 기반 방법[2,3,4]과 의미 태그 말뭉치를 구축하고 이를 학습하여 문제를 해결하는 지도 학습(Supervised Learning) 방법[1,5], 원시 말뭉치를 사용하여 단어의 의미를 구분하는 비지도 학습(Unsupervised Learning)방법[6] 등이 있다.

단어 공간 모델(Word Space Model)을 이용한 중의성 해소 방법은 의미 태그 말뭉치를 학습에 이용하는 지도 학습 방법의 하나로, 단어의 의미를 앞뒤에 나타난 또 다른 단어들의 벡터로 표현하여 각 단어의 의미 유사도

를 벡터 유사도로 대체하여 측정할 수 있는 모델이다[7]. 이 모델은 단어가 특정한 의미로 쓰였을 때, 함께 쓰인 다른 단어들로 표현이 가능하다는 가정을 한다.

[1] 연구에서는 한국어 단어 공간 모델을 구축하고 이를 기반으로 한 단어 중의성 해소 방법을 제안하였다. 단어 공간 모델은 의미 태그 말뭉치를 학습하여, 임의의 단어(w)에 태깅된 의미(s)를 문맥 내에 나타난 단어(v)들의 벡터로 나타내어 추출한다. 이후, 새로운 단어(q)가 단어 벡터 형태로 주어지면 이와 가장 유사한 의미 벡터를 계산하여 그 단어의 의미를 결정한다. w에 대해 m개의 동형이의어가 있고, 각 동형이의어 s_i가 문맥 단어들 v_i을 각각 n개 가지고 있다면 아래와 같이 식(1), 식(2), 식(3)으로 각각 표현된다. 여기에서 f_{ijk}는 s_i로부터 윈도우내의 거리 k만큼 떨어진 문맥 단어 j의 빈도를 뜻하며, d_{ijk}는 그때의 거리 계산 값을 뜻한다. d는 중심 단어로부터의 거리와 반비례하며 k윈도우일 경우, 가장 가까운 좌우 단어를 k값으로 시작하여 1씩 감소시켜 가장 먼 단어가 1값을 갖도록 한다. 이 연구에서는 k를 5로 정하여 계산했다. 단어 중의성 해소 문제는 식(4)로 표현되며, 각 인수는 식(5), 식(6)으로 계산된다.

$$w = \{s_1, s_2 \cdots, s_m\} \tag{1}$$

$$s_i = < v_{i1}, v_{i2} \cdots, v_{in} > \tag{2}$$

$$where\ v_{ij} = \frac{\sum_k f_{ijk} \times d_{ijk}}{\sum_j \sum_k (f_{ijk} \times d_{ijk})},\ f = \text{빈도}, d = \text{거리가중치}$$

$$q = < v_{q1}, v_{q2} \cdots, v_{qn} > \tag{3}$$

$$argmax_i\ Pr_{prior}(s_i) \times Sim(s_i, q) \tag{4}$$

$$Pr_{prior}(s_i) = \frac{freq(s_i)}{\sum_k freq(s_k)} \tag{5}$$

$$Sim(s_i, q) \equiv cos(s_i, q) = \frac{\sum_k (v_{ik} \times v_{qk})}{\sqrt{\sum_k (v_{ik} \times v_{ik})} \times \sqrt{\sum_k (v_{qk} \times v_{qk})}} \tag{6}$$

이 모델은 높은 정확도를 보이지만, 의미 단어 벡터의 차원 수가 커져 계산 시간이 오래 걸리는 단점이 있다. 이러한 단점을 보완하기 위해 본 연구에서는 Word2Vec를 이용하여 의미 단어 벡터의 차원을 축소한다[8].

3. Word2Vec를 이용한 단어 벡터 차원 축소

Word2Vec는 인공 신경망의 하나의 언어 모델로, 비지도 학습 방법을 사용하며 신경망 중간의 은닉계층에 표현된 내용이 차원 축소된 벡터로 이용된다. 또, 이는 입력 단어와 출력 단어의 선택 방법에 따라 CBOW (Continuous Bag Of Word) 모델 또는 Skip-gram 모델로 구분되는데, CBOW는 문맥이 주어지면 그 문맥에 적합한 단어가 출력되고, Skip-gram은 단어가 주어지면 그에 따르는 문맥이 출력된다. 본 연구에서는 단어의 문맥을 입력으로 하고, 그 단어가 그 문맥에서 가진 하나의 의미를 찾기 위한 것이므로 CBOW 모델을 이용한다.

말뭉치에서 쓰인 의미 단어를 의미 벡터로 변환하기 위하여 먼저 Word2Vec을 이용하여 모든 단어의 벡터를 추출한다. 단어 추출 시 말뭉치에서 나온 모든 단어가 아닌 실질적으로 영향력이 있는 명사, 형용사, 동사 단어만을 추출하였다[9].

Word2Vec에서 차원 축소된 단어의 고유벡터를 추출한 후 한국어 단어 공간 모델의 의미 단어 벡터를 형성한다. 기존 한국어 단어 공간 모델에서의 의미 단어 벡터와 같이 좌우 문맥의 단어가 의미 단어 벡터의 요소가 되며, 빈도와 거리 가중치를 두어 의미 단어 벡터를 생성한다. 이는 기존 모델의 식(2)를 변형한 것으로 식(7)로 표현된다. 여기에서 W2V (Word2Vec) 수행결과가 차원 축소된 벡터이므로 이 벡터들의 합인 의미 단어 벡터도 차원이 축소된다. 또, 축소된 벡터는 코사인 유사도 계산시 음수가 나올 수 있으므로 식(8)로 양수화한다.

$$s_i^{w2v} = \sum_{j=1}^{n} (v_{ij} \times W2V(wordform(v_{ij}))) \tag{7}$$

$$where\ v_{ij} = \frac{\sum_k f_{ijk} \times d_{ijk}}{\sum_j \sum_k (f_{ijk} \times d_{ijk})},\ f = \text{빈도}, d = \text{거리가중치}$$

$$Sim(s_i^{w2v}, q) \equiv \frac{\cos(s_i^{w2v}, q) + 1}{2} \tag{8}$$

4. 실험 및 평가

본 연구는 세종 의미 형태 분석 말뭉치[10]를 사용하여 90%를 학습하고 10%를 테스트하였다. 테스트에 사용된 데이터 구성은 표 1과 같다.

문장 개수	단어 개수	의미 단어 개수
89,994개	2,422,340개	442,914개

표 1. 테스트 데이터의 구성

컴퓨터의 실험 환경으로 운영체제는 Windows 8.1, CPU는 AMD FX(tm)-6300 Six-Core Processor 3.50GHz, RAM은 8GB의 DDR3를 사용하였다. Word2Vec는 50차원으로 하였고, 이 실험 결과를 기존 연구와 비교하여 표 2, 표 3에 나타냈다.

	기존연구 [1]	기존연구[1] (상위50단어)	본 연구 (50차원)
정확도(%)	94.02	90.64	93.44
포스팅 로드시간(sec)	31.4	5.84	18.99
수행시간(sec)	0.35	0.21	0.046

표 2. [1]연구의 테스트 데이터 실험 결과 및 비교

	기존연구 [1]	기존연구[1] (상위50단어)	본 연구 (50차원)
정확도(%)	93.99	90.07	93.32
포스팅 로드시간(sec)	32.3	6.1	19.0
수행시간(sec)	140.7	81.5	18.5

표 3. 본 연구의 테스트 데이터 실험 결과 및 비교

성능은 기존 모델[1]에서 성능이 가장 좋은 빈도 거리 가중치 단어 벡터를 기준으로 비교하였다. 또한, 속도 비교를 위해 기존 모델의 의미 벡터 포스팅 단어 벡터를 상위 50단어만 제한 생성하여 비교하였다. 표 2는 [1]연구에서 사용된 200문장의 테스트 데이터 셋으로 비교한 실험 결과이며 표 3은 본 연구의 테스트 데이터를 실험한 결과이다.

실험 결과, 제안한 방법이 기존 방법에 비해 정확도는 약 0.6% 정도 약간의 성능 하락이 있었지만, 수행 시간이 약 7.6배 정도 향상되었다. 또, 기존 방법의 포스팅 단어 벡터를 상위 50 단어로 제한하더라도, 본 연구 방법의 수행 속도가 여전히 4.4배정도 빠름을 알 수 있다.

5. 결론

단어 의미 모호성 해소는 문맥 정보를 이용하여 중의적 단어의 의미를 결정하는 것으로 사전이나 말뭉치 등을 이용하여 연구되어왔다. 단어 공간 모델을 이용한 방법은 말뭉치에 기반을 둔 것으로 많은 단어 처리 때문에 공간 모델의 차원이 매우 높아 속도가 느린 편이다. 본 논문에서는 기존의 단어 공간 모델의 차원을 축소하여 의미 모호성을 해소하는 방법을 제안하였다. 즉, Word2Vec를 이용하여 차원 축소된 문맥 벡터를 추출한 후 이를 단어 공간 모델에 적용하였다. 세종 형태 의미 분석 말뭉치에 대해 실험한 결과, 제안한 방법이 기존의 성능을 어느 정도 유지하면서도 수행속도가 매우 향상되었음을 보였다.

감사의 글

* 본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [R0101-15-0062, 휴먼 지식증강 서비스를 위한 지능진화형 WiseQA 플랫폼 기술 개발]

* 이 논문은 2014년도 충북대학교 학술연구지원사업의 연구비 지원에 의하여 연구되었음(This work was supported by the research grant of Chungbuk National University in 2014)

참고문헌

[1] 박용민, 이재성, “한국어 단어 공간 모델을 이용한 단어 의미 중의성 해소”, 한국콘텐츠학회논문지, 제12권, 제6호, pp. 41-47, 2012.6

[2] 허정, 옥철영, “사전의 뜻풀이말에서 추출한 의미 정보에 기반한 동형이의어 중의성 해결 시스템”,정보과학회논문지 소프트웨어 및 응용, 제28권,제9호, pp.688-698, 2001.

[3] 허정, 서희철, 장명길, “상호정보량과 복합명사 의미사전에 기반한 동음이의어 중의성 해소”, 정보과학회논문지 소프트웨어 및 응용, 제33권, 제12호, pp.1073-1089, 2006.
[4] 김준수, 최호섭, 옥철영, “가중치를 이용한 통계기반 한국어 동형이의어 분별 모델”, 정보과학회논문지 소프트웨어 및 응용, 제30권, 제11·12호, pp.1112-1123, 2003.
[5] 이호,백대호,임해창, “분류 정보를 이용한 단어 의미 중의성 해결”, 정보과학회논문지(B), 제 24권, 제7호, pp.779-789, 1997.
[6] 박성배, 장병탁, 김영택. "의미 부착이 없는 데이터로부터의 학습을 통한 의미 중의성 해소." 한국정보과학회 2000 년도 봄 학술발표논문집 27.1B pp.330-332, 2000.
[7] H. Schutze, “Automatic Word Sense Discrimination,” Computational Linguistics, Vol.24, No.1, 1998.
[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean ,“Efficient Estimation of Word Representations in Vector Space”, In Proceedings of Workshop at ICLR, 2013.
[9] Manning, D. Christopher and Schutze, Hinrich, Foundations of Statistical Natural Language Processing, MIT Press, pp.229-261, 1999.
[10] 국립국어원, 21세기 세종계획 최종성과물(2011년 12월 수정판), 2011.