단어 임베딩을 이용한 단위성 의존명사 분별[1)]

이주상○, 옥철영
울산대학교, 한국어처리연구실
dosa510@naver.com, okcy@ulsan.ac.kr

Disambiguation of Counting Unit Noun using Word Embedding

Ju-Sang Lee○, Cheol-Young Ock
Korean Language Processing Lab, University of Ulsan, Korea

요 약

단위성 의존명사는 수나 분량 따위를 나타내는 의존명사로 혼자 사용할 수 없으며 수사나 수관형사와 함께 사용하는 의존명사이다. 단위성 의존명사가 2가지 이상인 동형이의어의 경우 기존의 인접 어절을 이용한 동형이의어 분별 모델에서는 동형이의어 분별에 어려움이 있다. 본 논문에서는 단위성 의존명사 분별을 위해 단어 임베딩을 사용했으며 총 115,767개의 단어를 벡터로 표현하였으며 분별할 의존명사 주변에 등장한 명사들과의 유사도를 계산하여 단위성 의존명사를 분별하였다. 단어 임베딩을 이용한 단위성 의존명사 분별이 효과가 있음을 보았다.

주제어: 단위성 의존명사, 단어 임베딩, 동형이의어 분별, word2vec

1. 서론

의존명사는 홀로 쓰이지 못하는 비자립적 명사이다. 의존명사를 분별하기 위해서는 의존명사와 함께 사용한 단어에 영향을 받는다. 그 중 단위성 의존명사는 “자동차 한 대”, “어선 한 척”에서 ‘대’, ‘척’과 같은 수나 분량 따위를 나타내는 의존명사이다. 단위성 의존명사는 혼자 사용할 수 없으며 수사나 수관형사를 함께 사용하게 된다.

단위성 의존명사가 아닌 일반 의존명사는 바로 앞에 쓰이는 어절이나 함께 쓰인 단어의 정보를 이용하면 쉽게 분별이 가능하다. 하지만 단위성 의존명사가 동형이의어일 경우 이를 분별하기 위해서는 단위성 의존명사가 가리키는 사물의 명칭을 찾는 것이 중요하다. 예를 들어 “주사 두 대”와 “자동차 두 대”에서의 ‘대’는 동형이의어로, “주사 두 대”의 ‘대__01’는 “담배/매/주사를 피우는/때리는/놓는 횟수를 세는 단위”를, “자동차 두 대”의 ‘대__15’는 “차/기계/악기 따위를 세는 단위”로 다르게 사용된다. 이렇게 단위성 의존명사가 동형이의어인 경우 이를 분별하기 위한 사물의 명칭은 수사나 수관형사 앞에 위치하여 두 어절 이상 떨어지거나, 단위성 의존명사 뒤에 사용되는 경우가 있다. 만약 동형이의어 분별을 인접 어절만 사용하게 되면 위의 예제에서 ‘대’를 구분하기 위해 ‘두’라는 관형사 정보만을 사용하게 되므로 ‘대’를 올바로 구분하기 어렵다. 표준국어대사전에 등재된 의존명사 중 34종이 동형이의어이며, 품사가 의존명사가 아니지만 뜻풀이가 “~세는 단위”인 ‘단위성’ 명사(예, 가락)까지 포함하면 60종이 동형이의어이다.

본 논문에서는 단위성 의존명사가 동형이의어인 경우 이를 분별을 위해 단어 간의 벡터 유사도를 사용한다. 단어 벡터는 Word2Vec[1]의 Skip-Gram과 Negative Sampling[2]을 사용하여 말뭉치에 등장한 단어들을 50차원의 벡터로 구축한다. 그리고 단위성 의존명사 분별이 필요한 의존명사와 주변에 등장하는 명사들 간의 코사인 유사도(cosine similarity)를 구하여 단위성 의존명사 분별을 한다.

2. 관련 연구

최근 동형이의어 분별은 기존 말뭉치 기반의 분별 방법에서 말뭉치와 어휘 지식을 함께 사용하여 동형이의어 분별하는 방법이 연구되고 있다[3]. 이 방법은 한국어 어휘의미망(UWordMap)에서 단어의 상위어와 기존 말뭉치 기반의 알고리즘을 조합하여 동형이의어 분별을 한다. 단어를 분별하기 위해 해당 단어의 상위어들이 뒤에 오는 단어와 직접적으로 연관성을 보유하고 있는지 검색하여 동형이의어 분별을 하는 방법이다. 그리고 Word2Vec를 이용해 명사, 동사, 형용사 단어에 대한 벡터를 구축하여 그래프 기반 단어 중의성을 해소하는 방법도 연구되었다[4]. 단어 임베딩을 활용한 그래프 기반 단어 중의성 해소는 문장에서 중의성을 가지는 단어와 등장한 단어들 간의 코사인 유사도를 계산하여 중의성을 해소한다. 비지도 학습이 아닌 지도 학습으로 단어 임베딩을 사용하여 단어의 중의성을 해소한 연구도 있다[5].

단어 임베딩은 동형이의어 분별뿐만 아니라 여러 자연어 처리 분야에서 사용한다. 단어 임베딩은 텍스트로 표현된 단어를 숫자로 표현하는 방법이다. 일반적인 단어 임베딩은 기계학습에서 텍스트의 단어를 입력하기 위해서 사용하였다. 기계학습의 입력으로 단어 임베딩을 사용하기 때문에 단어 벡터의 결과에 따라 기계학습의 성

1) 이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (R0101-16-0176, Symbolic Approach 기반 인간모사형 자가 학습 지능 원천 기술 개발)

능에 중요한 영향을 준다. 딥 러닝(Deep Learning)을 사용한 개체명 인식 연구[6]와 한국어 의존 구문 분석 연구[7]에서 단어 임베딩을 사용하여 기계학습의 입력에 사용한다. 최근에는 단순히 기계학습의 입력으로 사용하는 것이 아니라 단어 임베딩의 특징인 유사한 위치에 사용되거나 의미가 유사한 단어들이 유사한 벡터를 가지는 것을 활용하여 자연어 처리 문제를 해결하는 연구[4,5]도 있다.

3. 단어 임베딩을 이용한 단위성 의존명사 분별

3.1 단어 임베딩

단어 임베딩은 단어의 의미가 비슷한 단어끼리 유사한 벡터 공간에 위치한다. 이러한 단어 임베딩의 특성을 이용하여 단위성 의존명사 분별에 사용한다. 단위성 의존명사 분별을 위해 Word2Vec를 사용하여 단어를 벡터로 표현하였다. Word2Vec는 기존 단어 임베딩 방식인 Neural Network Language Model의 느린 학습 속도를 개선한 방법이다. Word2Vec는 CBOW(Continuous Bag-of-Word)와 Skip-Gram 방식이 있다. 두 방식 모두 문장에서 학습할 단어의 주변 단어에 등장하는 단어들을 학습 데이터로 사용한다.

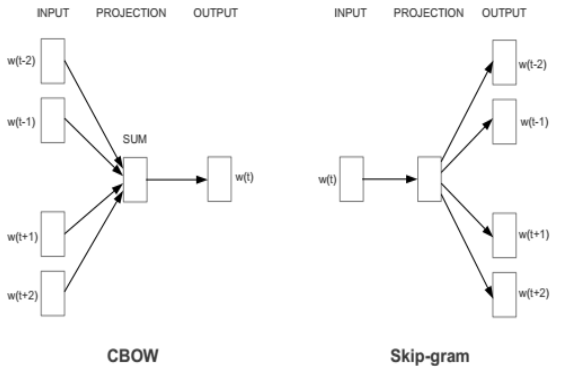

그림 1 Word2Vec의 CBOW(Continuos Bag-of-Word)와 Skip-gram

그림 1은 Word2Vec의 CBOW와 Skip-Gram을 그림으로 나타낸 것이다. 입력 값과 출력 값은 각 단어의 One-hot 형태로 학습하게 되며 Projection 층의 크기에 따라 하나의 단어가 가지는 벡터의 차원 수가 결정되게 된다.

Negative Sampling은 학습 데이터 이외에 임의의 부정적 데이터를 생성하여 현재 단어를 학습하는데 사용하는 방식으로 기존에 Output 층을 전부 계산하는 것이 아닌 일부만 계산하여 학습 속도가 빠르며 성능에 차이가 없다.

3.2 단위성 의존명사 분별

단위성 의존명사 분별을 위해서는 사물의 명칭이 필요하다. 하지만 사물의 명칭은 단위성 의존명사가 존재하는 어절의 바로 앞 어절과 뒤 어절에 존재 하지 않는 경우가 많다.

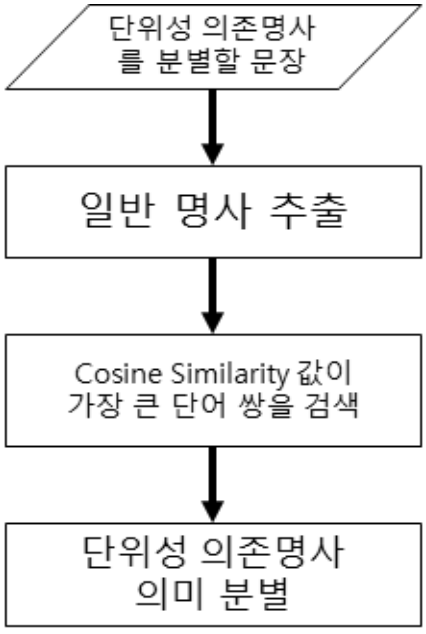

그림 2 단위성 의존명사 분별 흐름도

그림 2에서는 본 논문에서 사용한 단위성 의존명사 분별 방법을 나타낸다. 먼저 분별할 문장에서 단위성 의존명사를 포함하는 어절에 가까운 위치에 있는 명사들을 왼쪽과 오른쪽에서 두 개씩 추출한다. 추출한 일반명사들과 분별할 의존명사가 가질 수 있는 단위성 의존명사들을 코사인 유사도를 이용하여 가장 높은 유사도를 가지는 조합을 찾아 단위성 의존명사를 분별한다.

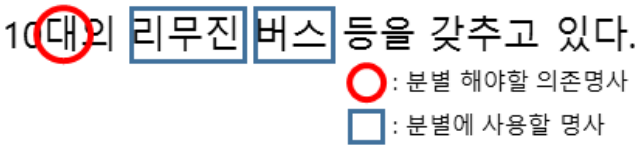

그림 3 단위성 의존명사 분별 예

그림 3는 단위성 의존명사 분별을 위한 예시이다. 사전에 등재된 '대'는 세 가지의 단위성 의존명사와 하나의 일반 의존명사를 가지고 있다. 먼저 '대__01'은 "때리거나 주사를 놓는 횟수"와 "담배를 피우는 횟수를 세는 단위"의 의미를 가지고 있으며 '대__11'은 "두 짝이 한 벌이 되는 물건을 세는 단위", '대__15'는 "차나 기계 악기 따위를 세는 단위"라는 의미를 가지고 있다. 3가지 단위성 의존명사 중에서 그림 3의 문장에 등장하는 '대'를 분별하기 위해 뒤에 근접해있는 일반명사인 "리무진"과 "버스"의 벡터를 사용하게 된다. 두 명사와 '대'의 세 가지 단위성 의존명사 중에 코사인 유사도를 사용하여 계산하게 되며 유사도 값이 가장 높은 값에 사용된 '대'의 의미로 결정한다.

4. 실험

본 논문에서는 실험을 위해 Word2Vec의 Skip-gram과 Negative Sampling을 사용하여 단어를 벡터로 표현한다. 학습 데이터로는 4,000만 어절로 구성된 말뭉치를 사용하였다. 동형이의어 수준으로 총 115,767개의 단어를 벡터로 표현하였다. 실험을 위해 단어 임베딩에 사용하지 않은 동형이의어 분별이 완료된 89만 문장에서 많이 등장한 단위성 의존명사에 대해 실험한다.

표 1 단위성 의존명사 정확률 실험

단어	단어의 빈도(%)	결과(%)
대__15	28.6	82.8
대__01	5.8	95.9
세__13	76.7	82.3
편__09	44.4	88.1
척__08	53.8	88.0

표 1은 5개의 단위성 의존명사에 대해서 앞과 뒤에서 두 개씩 명사를 추출해 유사도를 통한 분별을 실험한 결과이다. 단어의 빈도는 '대'라는 단어가 의존명사로 사용된 전체 빈도에서 '대__15'(차나 기계, 악기 따위를 세는 단위)가 등장한 빈도를 나타낸다. 실험 결과는 89만 문장에서 '대__15'와 주변에 등장하는 단어들이 모두 벡터 값을 가지는 문장만 사용하여 실험하였다. 빈도가 낮을수록 좋은 결과를 보여준다. 하지만 '대__15'의 경우 구별을 위한 사물의 명칭이 생략되어 앞 문장에 등장하거나 멀리 떨어져 위치하는 경우가 많아 빈도에 비해 결과가 낮게 나타났다.

표 2 명사의 수에 따른 단위성 의존명사 분별

단어	두 개 명사(%)	세 개 명사(%)
대__15	82.8	83.5
대__01	95.9	93.3
세__13	82.3	83.8
편__09	88.1	88.2
척__08	88.0	87.9

표 2은 분별할 의존명사에서 앞과 뒤로 명사의 수를 늘렸을 때의 결과이다. 명사의 수가 늘어나면 성능이 떨어지거나 늘어나는 명사가 있는 것을 볼 수 있다. 이를 통해 단위성 의존명사 분별을 위해서는 주변의 명사가 아닌 분별을 위한 사물의 명칭을 찾는 것이 중요하다는 사실을 보여준다.

5. 결론

본 논문에서는 단어 임베딩 중에 Word2Vec를 이용하여 단어 벡터를 구축하여 문장에서 분별할 의존명사의 주변에 등장하는 명사들을 이용하여 단위성 의존명사를 분별한다. 기존 모델에서는 단위성 의존명사 앞에 수사나 관형사가 쓰이기 때문에 분별의 어려움이 있었다. 하지만 단어 벡터를 이용한 단위성 의존명사 분별 실험을 통해 분별이 가능한 것을 볼 수 있었다. 단어의 빈도가 낮아도 분별력을 가지는 것을 실험을 통해 알 수 있었다. 하지만 분별을 위해 사용된 주변 명사에 해당 단위성 의존명사를 분별할 사물의 명칭이 없으면 분별력이 떨어지며 여러 명사 단어를 사용한 경우 사물의 명칭이 존재하여도 다른 의존명사와 유사도가 높아 오답이 나오는 경우도 있다. 또한 분별을 위한 사물의 명칭이 생략된 경우나 이전 문장에서 등장하는 경우 결과가 틀리게 된다.

향후에는 단위성 의존명사 분별을 위해 사물의 명칭을 찾아 분별하는 모델과 단위성 의존명사 분별을 위해 사용된 단어 임베딩을 개선할 모델을 연구할 계획이다.

참고문헌

[1] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed Representations of Words and Phrases and their Compositionality, NIPS 2013

[2] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119). 2013.

[3] 신준철, 옥철영. 한국어 어휘의미망(UWordMap)을 이용한 동형이의어 분별 개선. 정보과학회논문지. 제 43권. 제 1호. 71-79. 2016

[4] 오동석, 강상우, 서정연. Word2Vec을 이용한 반복적 접근 방식의 그래프 기반 단어 중의성 해소. 인지과학. 27(1). 43-60. 2016

[5] 신준철, 옥철영. 어휘지도(UWordMap)와 워드임베딩을 이용한 동형이의어 의미분별. 2016한국컴퓨터종합학술대회. 702-704. 2016

[6] 이창기, 김준석, 김정희, 김현기, "딥 러닝을 이용한 개체명 인식", 한국정보과학회 동계학술발표회 논문집, pp.423-425, 2014

[7] 이창기, 김준석, 김정희 "딥 러닝을 이용한 한국어 의존 구문 분석", 한글 및 한국어 정보처리 학술대회, 2014