용언의 의미 제약을 이용한 단어 임베딩[1)]

이주상[0], 옥철영
울산대학교, 한국어 처리 연구실
dosa510@naver.com, okcy@ulsan.ac.kr

Word Embedding using Semantic Restriction of Predicate

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

요 약

최근 자연어 처리 분야에서 딥 러닝이 많이 사용되고 있다. 자연어 처리에서 딥 러닝의 성능 향상을 위해 단어의 표현이 중요하다. 단어 임베딩은 단어 표현을 인공 신경망을 이용해 다차원 벡터로 표현한다. 본 논문에서는 word2vec의 Skip-gram과 negative-sampling을 이용하여 단어 임베딩 학습을 한다. 단어 임베딩 학습 데이터로 한국어 어휘지도 UWordMap의 용언의 필수논항 의미 제약 정보를 이용하여 구성했으며 250,183개의 단어 사전을 구축해 학습한다. 실험 결과로는 의미 제약 정보를 이용한 단어 임베딩이 유사성을 가진 단어들이 인접해 있음을 보인다.

주제어: Word embedding, Deep learning, 용언 의미 제약 정보, 한국어 어휘지도, UWordMap

1. 서론

최근 딥 러닝(Deep learning)을 이용한 한국어 자연어 처리 방법이 많이 사용되고 있다[1,2,3]. 딥 러닝을 자연어 처리에 사용하는 경우 Input layer의 입력 값으로 One-hot 형태([0100]처럼 해당 단어만을 1로 표현하는 binary 형태 방식)를 사용하면 구축한 단어 사전 수만큼 단어 벡터 차원이 높아지는 단점이 있다. 단어 벡터 차원이 높아지는 단점을 해결하기 위해서 딥 러닝을 이용한 자연어 처리에서는 단어 임베딩(word embedding)을 사용한다.

단어 임베딩은 단어들 간의 관계와 인공신경망(Neural Network)을 이용하여 단어를 다차원의 벡터 공간(Vector space)에 표현하는 방법이다. 단어 임베딩은 유사한 단어들이 벡터 공간상에 가깝게 위치하게 되어 기계학습에서 사전 학습의 효과가 있다. 또한 One-hot 형태를 사용하면 구축한 단어 사전의 크기만큼의 차원이 필요하지만 단어 임베딩을 통해 저차원 벡터 공간(low dimension Vector space)으로 차원 축소가 가능하다.

본 논문에서는 현재 한국어 단어 임베딩에서 사용되는 문장에서 현재 단어의 앞뒤 단어를 이용한 n-gram 방식이 아닌 용언의 목적격 필수논항 의미 제약 정보와 명사의 상하관계를 이용해 단어를 50차원의 벡터로 표현 하였다. 용언의 필수논항 의미 제약 정보와 명사의 상하관계 정보는 한국어 어휘지도인 UWordMap[4]을 사용하며 word2vec[5]의 Skip-gram과 negative-sampling을 이용한 단어 임베딩 방법을 제안한다.

1) 이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (R0101-15-0176)

2. 관련 연구

단어 임베딩은 단어 자질의 표현 방식으로 영어권을 시작으로 많은 연구가 있었다. 최근 딥 러닝이 화제로 떠오르면서 단어 임베딩에도 많은 연구가 있었다.

초기의 단어 임베딩은 대용량 원시 말뭉치(raw corpus)와 인공신경망을 이용해 문장에서 목표 단어와 목표 단어의 앞뒤 단어들을 이용해 목표 단어를 학습하는 n-gram 방식이 쓰였다[6]. 또한 인공신경망 학습 방법 중에 Recurrent Neural Network(RNN) 방식을 이용한 단어 임베딩인 Recurrent Neural Network Language Model(RNNLM)가 있다[7]. RNNLM은 이전 학습의 은닉층(hidden layer)의 결과 값을 다음 학습에 이용하는 방식이다. 학습의 속도를 높이기 위해 인공신경망의 은닉층(hidden layer)을 사용하지 않고 학습하는 word2vec가 있다[5].

현재의 단어 임베딩은 대용량 원시 말뭉치의 n-gram 방식의 접근이 아닌 의미적인 접근을 통한 단어 임베딩이 연구되고 있다. 현재 연구되고 있는 단어 임베딩 방법 중에는 문장에서 목표 단어의 의존관계(Dependency Relation)와 word2vec의 Skip-gram과 negative-sampling을 이용해 제안된 단어 임베딩 방법이 있다[8]. 벡터 공간에서 하위어를 찾는 방법 중에서 기존 n-gram 모델보다 의존관계나 의미적 접근이 더 좋은 성능을 보인다는 연구 결과도 있다[9]. 이를 통해 단순한 문장에서의 앞뒤 단어를 이용한 n-gram을 사용한 단어 임베딩이 아닌 의미적 접근이나 언어 구조적으로 접근하는 방법이 단어 임베딩에서 연구 되고 있다.

3. 용언의 의미 제약을 이용한 단어 임베딩

3.1 word2vec

본 논문에서는 단어 임베딩 학습 방법으로 word2vec를

사용했다. word2vec의 CBOW(Continuous Bag-of-Word)와 Skip-gram 중에서 Skip-gram을 사용했다. 그림 1은 CBOW와 Skip-gram을 그림으로 표현한 것으로 Skip-gram은 문장에서 목표 단어를 One-hot 형태의 표현 값을 Input layer의 입력 값으로 사용한다. 그리고 Output layer의 정답을 문장에서 목표 단어의 앞뒤 단어의 One-hot 형태의 표현 값을 사용한다. CBOW는 Input layer에 문장에서 목표 단어의 앞뒤 단어의 One-hot 형태의 표현 값을 사용하고 Output layer에 목표단어의 One-hot 형태의 표현 값을 사용한다. 또한 학습 결과를 향상시키기 위해 negative-sampling 기법을 이용한다. negative-sampling은 단어 임베딩 학습에서 학습 데이터 상의 정답 단어와 임의의 오답 단어를 함께 학습 하는 방법이다. 정답 단어는 Output layer에서 해당 단어만 1로 이루어진 One-hot 형태의 표현 값을 사용해 학습하고 오답 단어는 해당 단어를 0으로 설정한 One-hot 형태의 표현 값을 사용해 학습하게 된다.

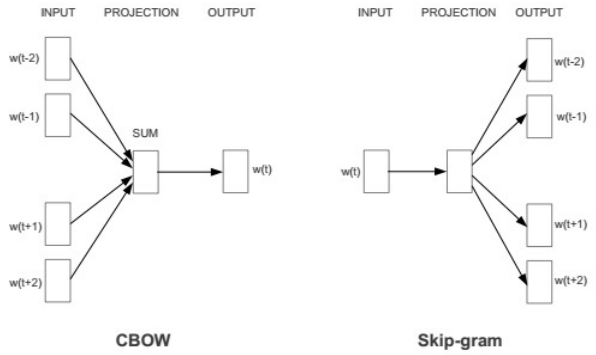

그림 1. word2vec의 CBOW와 Skip-gram

3.2 단어 임베딩 학습

본 논문에서는 한국어 어휘의미망으로 UWordMap[4]을 사용했다. UWordMap에 존재하는 명사와 용언을 동형이의어 수준으로 250,183개 단어 사전을 구축하고 타동사의 목적격에 대한 의미 제약 정보를 이용해 학습했다.

표 1은 용언 필수논항 의미 제약 정보를 보여주고 있다. 예를 들어 '감다__02'를 단어 임베딩에서 학습할 경우 Output Layer에는 '감다__02'의 목적어 의미 제약으로 올 수 있는 {신체부위, 가마, 눈동자, 동공} 등이 정답으로 사용된다. 명사를 학습하는 경우에는 해당 명사를 목적으로 취하는 용언을 Output Layer의 정답으로 사용해 학습하게 된다. 명사 학습에서 해당 명사를 의미 제약 정보로 가지고 있는 용언이 없는 경우 해당 명사의 상위어가 의미 제약 정보를 가지고 있으면 상위어가 가진 의미 제약 정보로 학습하게 된다. 예를 들어, '한국__05(대한제국을 줄여서 이르는 말, 대한민국)'의 경우는 '한국__05'를 목적어의 의미 제약으로 가지는 용언이 없다. 이 경우 '한국__05'의 상위어인 '제국__02(황제가 다스리는 나라)'와 '공화국__01(공화 정치를 하는 나라)'을 목적어의 의미 제약으로 가지는 용언을 추출하여 '한국__05'에 대해 단어 임베딩 학습을 한다.

표 1. UWordMap의 하위범주화 정보

용언(의미)	필수논항	명사
감다__02 (머리나 몸을 물로 씻다.)	을	신체부위, 가마, 눈동자, 동공, (후략)
갈다__01 (사람이나 물질을 바꾸다.)	을	부품, 물질, 칭호

4. 실험 결과

UWordMap을 통해 총 250,183개의 명사와 용언으로 구성된 단어 사전을 구축하고 단어 임베딩을 통해 50차원 벡터로 구성하였다. 그림 2와 그림 3은 t-distributed Stochastic Neighbor Embedding (t-SNE)[10]을 이용해 단어를 분포한 그림이다. 그림 2와 그림 3은 각각 명사 200개와 용언 200개를 임의로 추출해 나타낸 결과이다. 그림 2에서는 "산"을 뜻하는 단어들이 왼쪽에 모여 있다. 하지만 그림 3에서는 용언들은 명사에 비해 클러스터링 결과가 다른 모습을 보이고 있다.

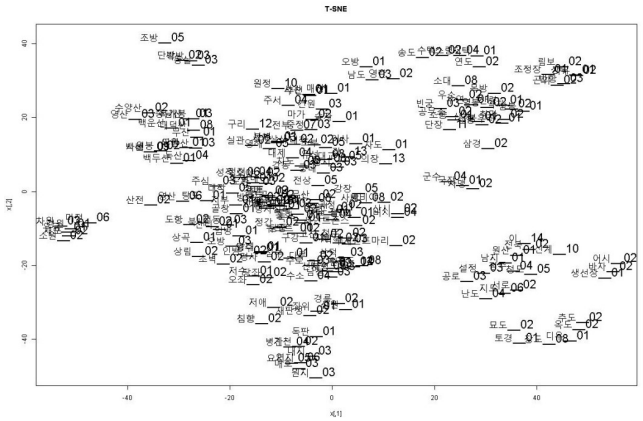

그림 2 명사 200개 분포(t-SNE 사용)

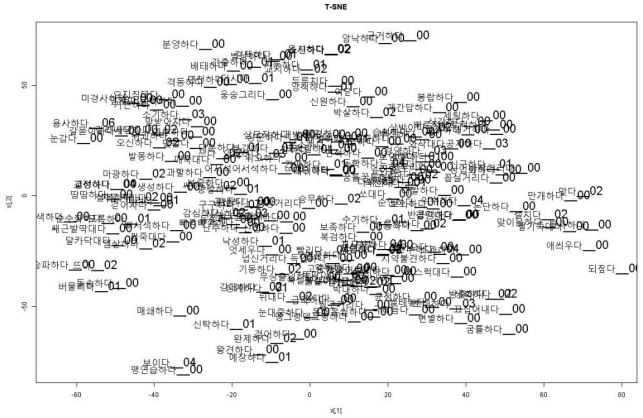

그림 3 용언 200개 분포(t-SNE 사용)

표 2는 각각의 명사에 대해 cosine similarity를 사용하여 인접한 명사 5개를 찾아낸 결과이다. 표 2의 결과를 통해 의미가 유사한 명사 단어들이 인접한 것을 볼 수 있다. 표 2에서 '미국__03' 의 경우에 국가 이름이나 국가 형태를 표현하는 단어가 나타났다. '중국__01' 의 경우 옛 국가 명칭과 국가 형태를 표현하는 단어가 인접해 나오는 결과를 볼 수 있다.

표 2. 3개의 명사와 가장 유사한 단어(괄호 안은 의미)

미국__03	설악산__00	중국__01
르완다	사덕산	바벨론
연방국	백사봉	정나라
참전국	학봉	중진국
합중국	청량산	가라(가야)
우크라이나	묘광산	번왕국

표 3은 용언에 대해 cosine similarity를 사용해 인접한 5개의 용언을 찾아낸 결과이다. 표 3의 결과를 통해 용언은 유사성이 조금 떨어짐을 보인다. '감다__02' 와 '서다__01' 은 의미적으로 유사한 용언들이 유사한 것이 아닌 해당 용언의 목적격 의미 제약 정보가 유사한 단어들이 나타났다. 반면, '걷다__02' 의 경우는 유사성이 부족하게 나타났다.

표 3. 3개의 용언과 가장 유사한 단어

걷다__02	감다__02	서다__01
무서워하다	맞부딪치다	내리갈기다
나서다	고정하다	비끄러매다
묻다__03	부들대다	놀다__01
내왕하다	붙들리다	구하다__01
돌다	부딪다	찔리다

5. 결론 및 향후 연구

본 논문은 UWordMap에서 용언의 목적격 의미 제약 정보와 word2vec[5]의 Skip-gram과 negative-sampling을 이용한 동형이의어 수준의 단어에 대한 단어 임베딩을 하였다. 명사에서는 유사성이 있는 단어들이 벡터 공간상에서 유사한 위치에 모여 있는 결과를 보였지만, 용언의 경우 명사에 비해 유사한 단어들이 모여 있지 않은 결과를 보였다. 하지만 단어의 의미 정보를 이용한 단어 임베딩이 효과가 있음을 보인다. 이는 기존의 단어 임베딩에서 사용되고 있는 문장에서 단어의 순서를 이용하는 n-gram 방식 이외에도 단어의 의미를 이용한 단어 임베딩이 효과가 있음을 보였다.

향후에는 명사의 상하 관계, 반의어, 용례 등의 다른 단어 의미적 요소를 이용해서 단어 임베딩을 적용할 계획이다. 또한 용언의 다른 필수논항 및 품사에 대해서도 단어 임베딩을 적용 해 볼 것이다.

참고문헌

[1] 나승훈, 정상근 "딥 러닝에 기반한 한국어 품사 태깅", 한국정보과학회 동계학술발표회 논문집, pp.426-428, 2014

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

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

[4] 옥철영, 배영준, "한국어 어휘지도(UWordMap)와 API 소개", 제 26회 한글 및 한국어 정보처리 학술대회 논문집, pp.27-31, 2014

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

[6] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A Neural Probabilistic Language Model, Journal of Machine Learning Research 3, pp.1137-1155, 2003

[7] T. Mikolov, S. Kombrink, L. Burget, J. Honza, S. Khudanpur, Extensions of Recurrent Neural Network Language Model, SLT, pp.234-239, 2012

[8] O. Levy, Y. Goldberg, Dependency-Based Word Embedding, 52nd Annual Meeting of the Association for Computational Linguistics Vol.2, 2014

[9] M. Rei, T. Briscoe, Looking for Hyponyms in Vector Space, 18th Conference on Computational Natural Language Learning, 2014

[10] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research, 9.2579-2605: 85, 2008