Word2Vec 모델을 활용한 한국어 문장 생성

남현규○, 이영석

충남대학교 컴퓨터공학과
hkanm@cnu.ac.kr, lee@cnu.ac.kr

Generating Korean Sentences Using Word2Vec

Hyun-Gyu Nam○, Young-Seok Lee
Chungnam National University, Dept. of Computer Engineering

요 약

고도화된 머신러닝과 딥러닝 기술은 영상처리, 자연어처리 등의 분야에서 많은 문제를 해결하고 있다. 특히 사용자가 입력한 문장을 분석하고 그에 따른 문장을 생성하는 자연어처리 기술은 기계 번역, 자동 요약, 자동 오류 수정 등에 널리 이용되고 있다. 딥러닝 기반의 자연어처리 기술은 학습을 위해 여러 계층의 신경망을 구성하여 단어 간 의존 관계와 문장 구조를 학습한다. 그러나 학습 과정에서의 계산양이 방대하여 모델을 구성하는데 시간과 비용이 많이 필요하다. 그러나 Word2Vec 모델은 신경망과 유사하게 학습하면서도 선형 구조를 가지고 있어 딥러닝 기반 자연어처리 기술에 비해 적은 시간 복잡도로 고차원의 단어 벡터를 계산할 수 있다. 따라서 본 논문에서는 Word2Vec 모델을 활용하여 한국어 문장을 생성하는 방법을 제시하였다. 본 논문에서는 지정된 문장 템플릿에 유사도가 높은 각 단어들을 적용하여 문장을 구성하는 Word2Vec 모델을 설계하였고, 서로 다른 학습 데이터로부터 생성된 문장을 평가하고 제안한 모델의 활용 방안을 제시하였다.

주제어: 문장 생성, 형태소 분석, word2vec

1. 서론

자연어처리는 인간의 언어 현상을 기계적으로 분석하여 컴퓨터가 이해할 수 있는 형태로 변환하고, 다시 인간이 이해할 수 있는 형태로 표현하는 기술을 의미한다. 최근 많은 기업들이 고도화된 머신러닝, 딥러닝 기술을 자연어처리 분야에 적용하여 서비스를 제공하고 있다. 구글 어시스턴스, 애플의 시리, 아마존 알렉사 등은 음성 인식 기술과 자연어처리 기술을 결합하여 사람의 발화를 문장화 하고 자연어처리를 통해 의도를 파악한다. 또한 채팅으로 사람과 대화할 수 있는 챗봇(Chatbot) 형태의 서비스 역시 자연어처리 기술을 적용하여 사람이 입력한 문장의 의미를 파악하고 사람이 이해할 수 있는 문장으로 결과물을 표현한다.

딥러닝 기반의 언어 모델은 여러 개의 문장과 단어, 말뭉치(Corpus) 데이터를 학습하여 단어와 문장 간의 의존관계를 분석한다. 문장의 일부가 주어졌을 때 나머지 부분을 추론하여 가장 높은 확률을 가진 단어들로 문장을 완성하는 확률 기반의 모델이며, 대표적으로 RNN과 LSTM 등이 있다. 딥러닝 기반 언어 모델은 학습 데이터로부터 문장의 특징을 자동적으로 학습하고 텍스트 뿐만 아니라 사진, 음성 등을 같이 활용할 수 있다. 그러나 실제 출력 계층의 차원이 크기 때문에 은닉 계층의 병렬 배치와 같이 연산 속도를 줄이는 작업이 추가로 필요하며, 작업을 뒷받침할만한 하드웨어 성능이 필요하다.

* 이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2016R1D1A1A09916326)

그러나 텍스트를 처리하기 위해 개발된 Word2Vec 모델은 신경망 구조를 유지하면서 선형 구조를 가지므로 이전 모델에 비해 적은 시간 복잡도로 단어 간 유사도를 계산할 수 있다. 단어를 벡터로 표현하여 학습 과정에서 한 단어를 기준으로 단어 주변의 문맥을 얼마나 정확하게 예측하는지 계산하고, 계산된 결과를 바탕으로 단어 간 관계를 파악할 수 있다.

따라서 본 논문에서는 Word2Vec 모델을 기반으로 하여 단어 벡터를 학습하고 한국어 문장을 생성하는 방법을 제안한다. 학습 대상이 되는 한국어 문장을 형태소 분석을 통해 품사별로 단어 벡터를 생성한다. 생성한 단어 벡터는 Word2Vec 모델에 적용하여 유사도가 높은 단어 벡터를 사전에 지정한 문장 템플릿의 주어, 목적어 등 문장의 각 구성 요소로 하여 한국어 문장을 생성하였다. 학습 데이터에 따라 생성된 문장이 어떻게 달라지는지 분석하기 위해 뉴스 기사와 온라인 커뮤니티 게시물을 수집하여 각각 학습 모델을 만들고, 동일한 주어로 시작하는 문장을 생성하여 서로 비교하였다.

2. 관련 연구

Word2Vec 모델은 단어 학습의 계산 복잡도를 최소화하기 위해 고안되었다[1]. Feed-Forward Neural Net Language Model(NNLM)[2]의 한계를 극복하기 위해 고안되었다. NNLM 모델은 단어의 특징 벡터를 학습하여 단어의 분포를 구하는 과정에서 투영 계층과 출력 계층 간 연산이 오래 걸리는 문제점이 있다. NNLM의 한계를 해결

하기 위해 고안된 Recurrent Neural Network Language Model(RNNLM)은 투영 계층 없이 입력-은닉-출력 계층으로 신경망을 구성하고, 은닉 계층에서는 모든 단어에 대한 확률 분포를 계산하여 시간 흐름에 따라 재귀적으로 반복 학습한다. RNNLM은 NNLM에 비해 상대적으로 시간 복잡도가 여전히 높다는 문제점이 있다. 그러나 Word2Vec 모델은 여러 개의 주변 단어를 통해 대상 단어를 유추하는 Continuous Bag-of-Word(CBOW), 주어진 단어 하나를 가지고 주위에 등장하는 나머지 단어들을 유추하는 Skip-gram 으로 구성되어 가공되지 않은 텍스트로부터 효율적으로 예측할 수 있다. 또한 해당 단어가 나올 확률을 예측하는 트리를 구성할 때 Binary Huffman Tree를 사용한다. 즉, 자주 등장하는 단어들이 보다 짧은 path로 도달하게 되어 전체적인 계산 복잡도가 낮아지는 효과가 있다.

[3]에서는 여러 문장을 만들어 둔 상태에서 상황과 맥락에 따라 문장을 끌어다 사용하는 방식으로 프로야구 경기를 요약한 기사를 작성하였다. 문장의 주어와 서술어 등을 비워 둔 상태에서 프로야구 경기 데이터를 결합하여 뉴스 기사의 각 문장을 생성하고, 기록 시점 이전에 경기 데이터와 현재 이벤트 이후 경기 데이터를 비교하여 '안타깝게도' 등과 같이 무드를 더하여 문장을 생성하였다. 그러나 이 연구는 문장 템플릿에 사용하는 목적어와 서술어가 정형화 되어 있고, 새로운 문장을 생성하기 위해서는 템플릿을 추가하여 경기 데이터를 적용해야 한다는 불편함이 있다. 본 연구에서 사용하는 Word2Vec 모델은 템플릿을 적용한 기본 문장을 구성한 후 워드 벡터에서 가장 유사한 품사들을 추가하여 문장을 확장할 수 있으므로 이벤트마다 별도의 문장 템플릿이 필요하지 않다.

3. 한국어 문장 생성 모델

3.1 형태소 분석

Word2Vec 모델에서 단어 벡터는 문장을 구성하는 각각의 품사가 된다. 따라서 Word2Vec 모델을 구성하기 위해서는 학습 데이터를 형태소 분석을 통해 문장을 어근, 접두사/접미사, 품사 등으로 가장 세분화하여 단어 벡터로 생성하는 전처리 과정이 필요하다. 본 연구에서는 파이썬 한글 형태소 분석 라이브러리인 KoNLPy[4]의 트위터 형태소 분석기를 이용하여 단어 벡터를 구성하였다. 그림 1은 학습할 한국어 문장을 형태소 분석한 예시이다. 분석 과정에서 품사(POS) 태그를 함께 태깅 하여 동음이의어를 구분하였다. 또한 학습 데이터 중 복합 명사의 경우 형태소 분석 과정에서 명사와 명사로 분리되지 않도록 미리 사전을 구성하여 하나의 명사로 분류하였다.

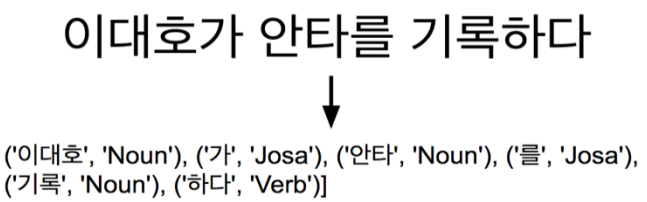

그림 1 형태소 분석 예시

3.2 Word2Vec 모델 구성

형태소 분석을 통해 품사 태그가 포함된 단어들로 벡터를 구성하여 Word2Vec 모델을 생성하였다. Word2Vec 모델은 학습 과정에서 기준이 되는 단어로부터 주변의 단어 문맥을 파악하여 현재의 단어 벡터 위치가 얼마나 정확한지를 계산한다. 예를 들어, 차원의 크기가 100인 벡터 공간에서 각 워드 벡터는 100차원 공간의 점 하나에 해당되며, 학습 과정을 통해 의미가 유사한 단어들이 근처에 위치하게 된다. 단어 간 유사도가 높다는 것은 벡터 공간에서 단어 간 거리가 가깝다는 것을 의미한다.

학습이 완료되면 미리 지정한 템플릿에 Word2Vec의 단어 벡터를 이용하여 문장을 생성한다. 주어와 유사도가 가장 높은 단어 벡터들이 목적어, 서술어 등 나머지 문장의 구성 요소가 되어 전체 문장을 완성한다. 그림 2는 단어 벡터들을 t-SNE 기법으로 시각화한 예시이다.

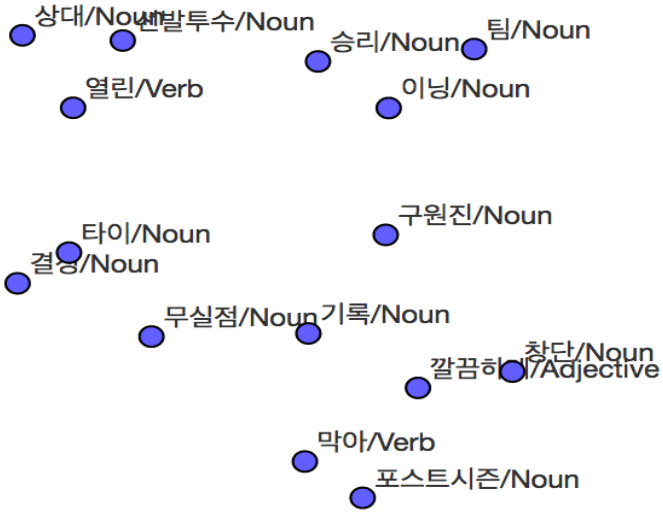

그림 2 Word2Vec 단어 임베딩 예시

벡터 공간에 단어가 임베딩이 완료되면 단어 벡터 간 유사도를 계산하여 결과값이 가장 높은 단어 벡터를 이용하여 문장을 생성한다. 먼저 생성할 문장의 주체가 되는 명사를 기준으로 가장 유사도가 높은 단어 벡터 중 조사와 명사를 찾는다. 조사는 명사와 함께 결합하여 '~은', '~가'와 같이 문장의 주어를 구성하고, 주어와 가장 유사도가 높은 명사는 목적어가 되어 주어가 하는 행위의 대상이 된다. 목적어 역시 주어와 마찬가지로 유사

도가 가장 높은 조사와 동사를 찾아 문장의 나머지 구성 요소인 목적어와 서술어로 사용한다. 그림 3은 Word2Vec 모델을 이용하여 문장을 생성하는 전체 과정을 플로우 차트로 나타내었다.

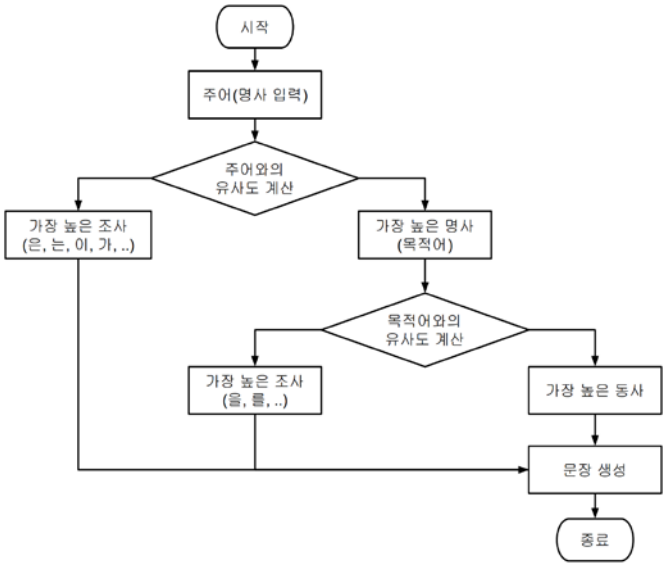

그림 3 유사도 기반 문장 생성 플로우차트

4. 한국어 문장 생성 예시

4.1 학습 데이터

본 논문에서 제안한 Word2Vec 모델의 문장 생성 성능을 평가하기 위해 동일한 주제의 두 가지 학습 데이터를 이용하였다. 학습 데이터로는 네이버 스포츠의 야구 뉴스 기사 1,575개의 본문과 대표 온라인 커뮤니티 디시인사이드의 국내야구 갤러리 게시물 38,724개를 수집하였다. 뉴스 기사와 커뮤니티 게시물 모두 같은 기간 동안 생성된 일주일 분 텍스트 데이터를 수집하였으며, 커뮤니티 게시물의 경우 뉴스와는 달리 본문보다는 제목에 의견을 표현하므로 학습할 문장의 개수를 조절하여 뉴스에 비해 상대적으로 많은 양의 게시물을 수집하였다.

4.2 실험 결과

생성한 문장의 템플릿은 지정한 주어와 유사도가 가장 높은 명사를 목적어로 하고, 목적어와 가장 유사도가 높은 동사를 서술어로 하여 (주어+목적어+서술어) 형태로 구성하였다. 주어와 목적어와 함께 사용되는 조사 역시 각각의 단어 벡터와 가장 유사한 조사를 사용하였다. 두 학습 모델을 서로 비교하기 위해 동일한 주어로 유사도를 계산하여 문장을 생성하였다. 표 1은 두 가지 데이터로 학습한 Word2Vec 모델이 생성된 문장의 예시이다.

표 1 문장 생성 결과

NO	뉴스 기사	커뮤니티 게시물
1	**롯데**가 승리를 하다	**롯데**가 극장을 하다
2	**최진행**이 삼진을 당하다	**최진행**이 진짜 이다
3	**한화**가 감독을 바꾸다	**한화**가 감독은 나가다
4	**이대호**는 도루를 하다	**이대호**는 돼지 이다
5	**한화**가 실책을 하다	**한화**는 수비가 망했다

학습 데이터에 따라 생성된 문장의 주어-서술어 간 관계가 어색할 수 있다. 문장 3의 뉴스 학습 데이터는 주어-서술어 구조를 유지한 상태에서 '바꾸다'라는 서술어로 일관성 있게 작성된 반면, 커뮤니티 게시물 학습 데이터의 대다수는 주어-서술어 구조가 아닌 '나가라/나가'와 같이 명령형으로만 작성되었다. 따라서 커뮤니티 게시물 기반 문장은 뉴스 기사에 비해 명사와 동사 간 유사도가 낮아 상대적으로 문맥상 어색한 부분이 있다. 또한 같은 의미를 가진 문장이지만 학습 데이터에 따라 서로 다른 표현으로 생성될 수 있다. 문장 1과 2는 커뮤니티에서 주로 사용하는 표현인 극적인 승리를 '극장'으로, '실책을 하다'라는 문장은 '진짜다'로 생성되었다. 문장 4와 5에서는 뉴스에서는 찾을 수 없는 개인의 주관적인 표현을 확인할 수 있다.

5. 결론

본 논문에서는 Word2Vec 모델을 이용하여 한국어 문장을 생성하는 방법에 대해 제안하였다. Word2Vec 모델은 충분한 양의 데이터로 학습할 경우 높은 정확도로 단어 간 유사도를 계산할 수 있었으며, 다른 텍스트 모델에 비해 상대적으로 계산 효율이 높음을 확인할 수 있었다. 그러나 단어 벡터를 구성하는 과정에서 형태소 분석이 제대로 이루어지지 않을 경우 전달하고자 하는 의미가 달라지거나 문맥상 어색한 문장이 생성될 수 있으며, 신경망 모델처럼 단어 간 의존관계를 파악하여 문장을 자동으로 생성하기에 어려움이 있다. 또한 학습 데이터에 따라 생성하는 문장의 표현이 서로 달라진다. 예를 들어, '실책' 이라는 이벤트에서 뉴스 기사 기반의 모델에서는 '실책을 하다'로 생성하였지만, 커뮤니티 게시물 기반 모델에서는 '수비가 망했다'로 표현하였다. 이것은 같이 사용되는 단어들이 유사도가 높게 측정되기 때문에 같은 주어를 사용하더라도 서로 다른 내용의 문장을 함께 학습할 경우 유사도가 높더라도 함께 사용된 품사가 무엇인가에 따라 달라질 수 있다.

그러나 유사도를 기반으로 하여 문장을 생성하여 주어와 목적어, 목적어와 서술어 간의 관계가 잘 표현되어 뉴스 기사와 같이 객관적인 사실을 전달하는 문장을 생성하거나 문서의 내용을 요약하는 문장을 생성할 경우에는 활용 가능성이 있음을 확인할 수 있었다.

참고문헌

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean.

Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.
[2] C. Chelba, M. Norouzi, and S. Bengio. N-gram language modeling using recurrent neural network estimation. arXiv preprint arXiv:1703.10724, 2017.
[3] 김동환, 이준환. (2015). 로봇 저널리즘 : 알고리즘을 통한 스포츠 기사 자동 생성에 관한 연구. 한국언론학보, 59(5), 64-95.
[4] 박은정, 조성준, “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지”, 제 26회 한글 및 한국어 정보처리 학술대회 논문집, 2014.