딥러닝 기반의 뉴스 분석을 활용한 주제별 최신 연관단어 추출 기법[1)]

김성진, 김건우, 이동호[2)]
한양대학교 컴퓨터공학과
e-mail: {tookdown, kgwhsy, dhlee72}@hanyang.ac.kr

A Topic Related Word Extraction Method Using Deep Learning Based News Analysis

Sung-Jin Kim, Gun-Woo Kim, Dong-Ho Lee
Dept of Computer Science, Hanyang University

요 약

최근 정보검색의 효율성을 위해 데이터를 분석하여 해당 데이터를 가장 잘 나타내는 연관단어를 추출 및 추천하는 연구가 활발히 이루어지고 있다. 현재 관련 연구들은 출현 빈도수를 사용하는 방법이나 LDA와 같은 기계학습 기법을 활용해 데이터를 분석하여 연관단어를 생성하는 방법을 제안하고 있다. 기계학습 기법은 결과 값을 찾는데 사용되는 특징들을 전문가가 직접 설계해야 하며 좋은 결과를 내는 적절한 특징을 찾을 때까지 많은 시간이 필요하다. 또한, 파라미터들을 직접 설정해야 하므로 많은 시간과 노력을 필요로 한다는 단점을 지닌다. 이러한 기계학습 기법의 단점을 극복하기 위해 인공신경망을 다층구조로 배치하여 데이터를 분석하는 딥러닝이 최근 각광받고 있다. 본 논문에서는 기존 기계학습 기법을 사용하는 연관단어 추출연구의 한계점을 극복하기 위해 딥러닝을 활용한다. 먼저, 인공신경망 기반 단어 벡터 생성기인 Word2Vec를 사용하여 다양한 텍스트 데이터들을 학습하고 룩업 테이블을 생성한다. 그 후, 생성된 룩업 테이블을 바탕으로 인공신경망의 한 종류인 합성곱 신경망을 활용하여 사용자가 입력한 주제어와 관련된 최근 뉴스데이터를 분석한 후, 주제별 최신 연관단어를 추출하는 시스템을 제안한다. 또한 제안한 시스템을 통해 생성된 연관단어의 정확률을 측정하여 성능을 평가하였다.

1. 서론

최근 SNS, 전자 메일, 검색 엔진 서비스 등 온라인상에서 많은 양의 텍스트 데이터가 생성되고 있다. 이러한 데이터들은 매우 빠르고 방대하게 생성되고 있기 때문에, 사용자가 단시간 내에 원하는 주제에 대한 정보를 파악하기가 어려워지고 있다.

최근 검색엔진들은 사용자가 원하는 주제와 관련된 연관단어들을 다른 사용자들이 입력한 검색어, 즉 연관 검색어들을 통해서 제공하고 있다. 따라서 특정 주제와 관련해서 다수의 사용자들이 최근 자주 사용하는 검색들을 쉽게 파악할 수 있다는 장점이 있다. 반면, 특정 주제와 관련해서 최근 주요 이슈와 관련된 정보를 효과적으로 제공 하는데 는 한계점이 존재한다. 예를 들어, 어떤 사용자가 특정 핸드폰 제조회사에 대해 최근 주요 이슈(즉, 해당 회사와 관련된 최근 정치, 경제, 사회, 문화적 이슈 등)와 관련된 키워드를 알고 싶다고 가정 할 때, 대다수의 사람들이 주로 제조사의 최신 핸드폰 모델만을 검색하기 때문에 기존 시스템들은 대부분 핸드폰 모델 관련 정보만 우선적으로 제공하고 최근 주요 이슈들은 제대로 제공하지 못한다는 한계점을 가지고 있다. 결국, 특정 주제와 관련된 최근 이슈들을 검색하기 위해서는 사용자가 이미 특정 주제와 관련된 최근 정치, 경제, 사회, 문화적 이슈를 미리 파악하고 있어야만 관련된 정보를 검색할 수 있다는 단점이 존재한다.

본 논문에서는 이러한 단점을 극복하기 위해 사용자가 원하는 주제어와 관련된 근래의 뉴스 데이터를 분석하고 최근 주목받고 있는 여러 키워드들을 연관단어로 추출하여 사용자에게 제공하는 시스템을 구축한다.

기존의 연관단어 관련 연구는 출현 빈도수를 사용한 모델부터 LDA 같은 기계학습 기법을 활용하는 다양한 연구가 진행되었다.[1][2] 이중 성능 적으로 좋은 결과를 보이는 LDA 기법은 다음과 같은 한계점을 지니고 있다. 먼저, 찾을 토픽의 개수가 미리 정해져 있어야 한다. 또한, 비 계층 구조로 이루어져 토픽간의 연관성을 찾아낼 수 없다는 한계점을 지니고 있다.

이러한 전통적인 기계학습 기법의 한계점들을 극복하

1) 이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2016R1D1A1A09918271, 다중 저장소 지속성 환경에서 빅데이터 기술을 활용한 개인 맞춤형 소셜 멀티미디어 태깅 및 태그 관리 시스템)
2) 교신저자

기 위해 최근 딥러닝에 대한 연구가 많이 이루어지고 있다.[3] 딥러닝은 인공 신경망을 다층구조로 배치하여 데이터를 분석하는 기계학습의 한 분야이다. 딥러닝은 전통적인 기계학습 기법과는 달리 좋은 결과를 얻기 위해 특징(Feature) 설계에 많이 의존하지 않는다. 원시 데이터에서 보다 단순한 특징부터 복잡한 특징까지 자동으로 감지하고 생성해 내며 여러 파라미터들을 학습과정을 통해 자동적으로 조정해 나가는 장점을 가진다. 또한, 최근 다양한 응용 분야에서 기존 방법들보다 좋은 성능을 보여준다고 보고되고 있다. 특히 자연어 처리, 비전, 음성인식, 패턴인식 등의 분야에서 좋은 결과를 얻고 있으며 텍스트 처리에도 최근 많은 연구들이 진행되고 있다 [4],[5],[6].

본 논문에서는 기존의 연관단어 추출에 사용되던 기계학습 기법에서 벗어나 딥러닝을 활용하여 연관단어를 추출한다. 먼저, 검색엔진에서 생성되는 다양한 텍스트 데이터를 인공 신경망 기반 단어 벡터 생성기인 Word2Vec[7]을 통해 단어의 벡터를 계속해서 학습시켜 룩업 테이블(Lookup Table)을 생성한다. Word2Vec는 단어가 말뭉치 내에서 가지는 의미와 역할을 고려하여 값을 측정한다. 생성된 룩업 테이블을 바탕으로 합성곱 신경망 (Convolutional Neural Network)을 사용하여 주제와 관련된 최근 뉴스기사들을 분석하여 특징벡터로 학습하고 기사의 대표적인 특징을 가장 잘 표현할 수 있는 단어를 연관단어로 추출한다. 합성곱 신경망은 인풋 데이터를 저차원의 벡터로 변환하고 중요한 특징 데이터를 보존하며 다른 신경망들보다 상대적으로 적은 파라미터를 가져 효율적으로 학습을 가능하게 하는 강점을 가지고 있다. 합성곱 신경망은 본래 컴퓨터 비전 분야에서 많이 쓰였지만, 최근 다양한 연구를 통해 자연어 처리 분야에도 좋은 성능을 보이며, 최근 텍스트 처리 연구에도 많이 활용되어 매우 좋은 성능을 보여주고 있다 [4],[5],[6]. 본 논문에서는 합성곱 신경망을 통해 문서의 특징 벡터를 학습하여 학습된 벡터를 토대로 연관단어를 찾아낸다. 따라서 특징 벡터를 더 잘 학습할 수 있는 특징을 가진 합성곱 신경망의 강점에 주목하여 합성곱 신경망을 통해 연관단어를 추출한다.

2. 관련 연구

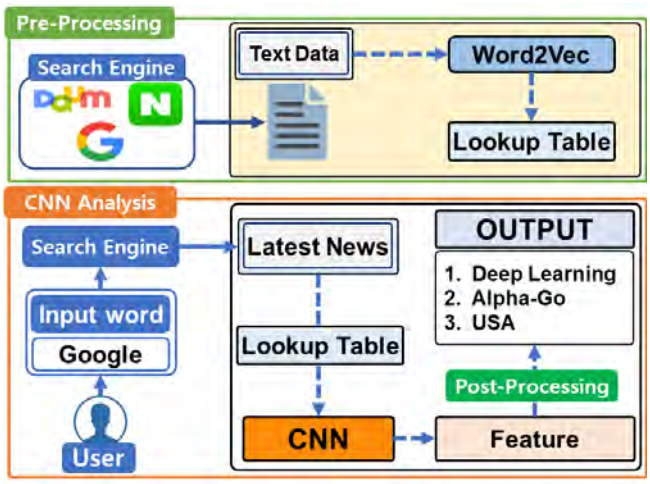
연관단어 연구는 출현 빈도수를 통해 연관단어를 생성하는 연구부터 기계학습기법인 LDA를 사용하여 연관단어를 생성하는 다양한 연구들이 진행 중에 있다 [1][2].

딥러닝은 주로 이미지 처리 분야에 많은 연구가 진행되었지만, 최근 텍스트 처리에도 딥러닝이 활용되면서 그 연구 범위를 넓히고 있다. 그 중, 합성곱 신경망을 통해 텍스트를 분석하는 연구들이 많이 진행되고 있다.

합성곱 신경망을 통한 단문 텍스트 랭킹, 합성곱 신경망을 통해 문장을 분석하여 분류하는 연구, 문장의 정서를 파악하는 연구 등 텍스트를 처리하는 다양한 연구에 합성곱 신경망이 활용되고 있다 [4],[5],[6].

하지만 현재까지 연관단어를 생성하는데 딥러닝을 사용한 연구는 없으며, 합성곱 신경망을 연관단어에 접목시킨 연구 또한 아직까지 진행되지 않았다.

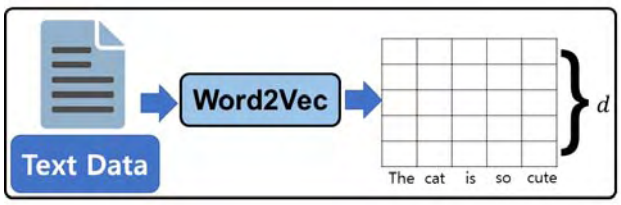
3. 제안 시스템



(그림 1) 전체 시스템 구조도

전체 시스템은 (그림 1)과 같다. 전체 시스템은 전처리 과정을 통해 다양한 검색엔진에서 실시간으로 생성되는 다양한 텍스트 데이터들을 Word2Vec를 통해 계속해서 학습한다. 학습을 통해 벡터 테이블이 만들어지면 인풋 주제와 관련된 최근 일주일간의 뉴스 기사들을 수집한다. 합성곱 신경망 분석과정에서는 생성된 벡터테이블을 룩업테이블로 사용하여 합성곱 신경망을 통해 수집한 뉴스 기사들의 단어들을 벡터로 변환하고, 합성곱 신경망을 통해 뉴스기사를 학습하여 주제어를 가장 잘 나타내는 특징 벡터를 학습한다. 마지막으로, 학습된 특징벡터가 의미하는 단어를 추출하여 연관단어를 생성한다.

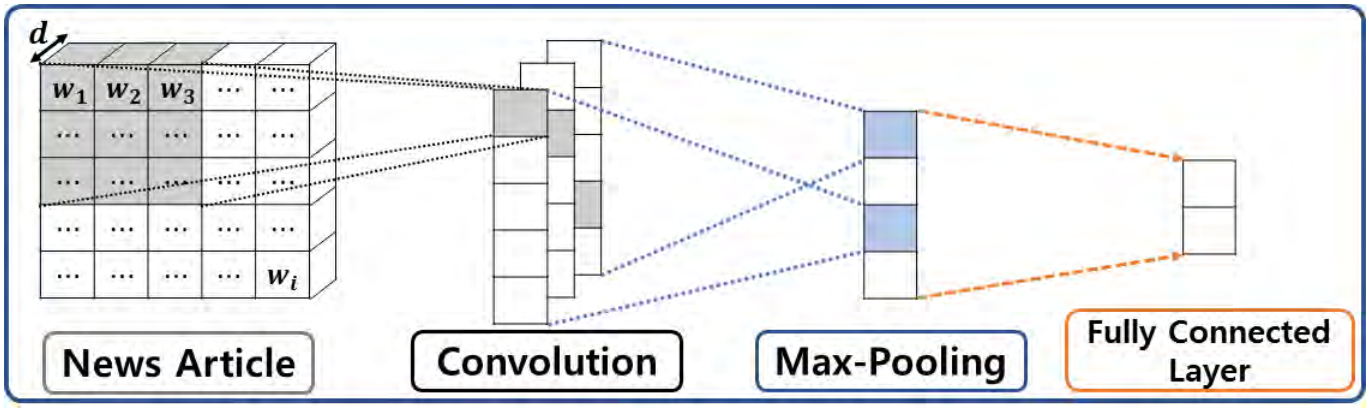
3.1 전처리 과정



(그림 2) 전처리 과정

전처리 과정에서는 합성곱 신경망을 통해 텍스트 데이터를 처리하기 위해 텍스트 데이터를 벡터 값으로 나타내는 과정을 거친다.

(그림 2)는 Word2vec를 통해 벡터 테이블을 생성하는 과정을 보여준다. 텍스트 데이터를 Word2Vec를 통해 분석하면, 단어마다 사용자가 사전에 지정한 d-차원의 벡터 값이 생성된다. 여러 검색엔진에서 실시간으로 생성되는 많은 텍스트 데이터들을 계속해서 Word2Vec를 통해 학습한다. 이를 통해, 단어를 벡터로 변환할 수 있는 벡터 테



(그림 3) 합성곱 신경망을 통한 분석 과정

이블을 생성하게 된다. 생성된 벡터 테이블은 합성곱 신경망 분석 과정에서 뉴스기사의 각 단어들을 벡터로 변환시키는데 사용된다.

3.2 합성곱 신경망을 통한 분석 과정

(그림 3)은 합성곱 신경망을 통해 뉴스기사를 분석하여 특징벡터를 학습하는 과정을 나타낸다. 전처리과정에서 학습과정이 계속해서 이루어지고, 유저가 입력한 주제어가 주어지면, 검색엔진에서 주제어와 관련된 일주일 분량의 뉴스기사들을 찾아낸다. 본 논문의 최종 목적은 주제어를 잘 나타내는 연관단어를 추출하는 것이기 때문에, 더 정확한 분석을 위해 뉴스기사의 단어들 중 be동사, 대명사, 관사 등 불용어를 제거하는 작업이 필요하다. Stanford POS Tagger 라이브러리를 사용해 단어의 품사를 판단하여 불용어를 제거하는 과정을 먼저 거친다. 불용어를 제거 한 이후, 뉴스기사(N)는 단어(w)의 조합으로 이루어지게 된다.

$$N = (w_1, w_2, w_3 \ldots w_n)$$

이후, 각 단어는 미리 생성된 룩업테이블을 통해 d-차원의 벡터 값으로 먼저 변형된다. 합성곱(Convolution) 과정은 다음과 같다. Stride값(s)을 정하고 뉴스기사에서 Stride 값만큼 단어(w)를 추출한다. Stride는 인풋 행렬에서 얼마만큼의 간격으로 필터를 적용할지를 정하는 값이다. 추출된 단어(w)와 필터의 계수(c)들을 순서대로 곱하여, 그 합을 피처 맵에 차례대로 채워 나간다. 즉, 합성곱 과정은 인풋 데이터의 각 부분들의 여러 값들을 하나의 값으로 표현한다. 필터의 계수(c)는 무작위의 값으로 초기화시킨 후, 학습과정을 진행하면서 학습해 나간다. 합성곱 과정은 다음과 같은 수식으로 표현 할 수 있다.

$$U = f(c \bullet w_n + b)$$

b는 각 필터마다 다르게 설정된 값인 바이어스 값이다. 활성화 함수인 f는 ReLU[8] 함수를 사용한다. 따라서 합성곱 과정에서는 Window 크기만큼 단어를 추출하여 위의 수식을 통해 Feature map(U)을 생성한다.

$$U^k = u_1^k, u_2^k, u_3^k, \ldots, u_i^k, k = Number of filter$$

이후, 전체 단어를 대상으로 Stride값만큼 이동시키며 반복적으로 Feature map(U)을 생성한다. 합성곱 과정을 거치고 난 후, 생성된 모든 Feature map(U^k)에 대하여 Max-Pooling과정을 수행한다.

$$m_k = \max U^k$$

k개의 Feature map(U^k)이 생성된 후, Max-Pooling은 각 Feature map의 값들 중 최댓값을 뽑아낸다. 생성된 Feature map의 개수(k)만큼 결과물(m_k)이 생성된다. 이 과정을 주제어와 관련된 최근 일주일 분량의 기사들을 순차적으로 입력받아 합성곱, ReLU 활성화 함수, Max-Pooling과정을 계속해서 원하는 만큼 수행 한 후, 전결합층(Fully Connected Layer)을 배치하여 학습을 진행한다. 이 과정을 반복하며 특징벡터를 학습한다.

3.3 후처리 과정

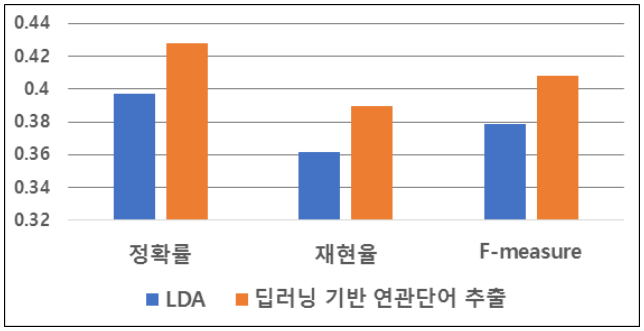
학습 과정을 마치면, 주제어와 관련된 최근의 뉴스기사들이 나타내는 특징들이 학습된다. 특징은 학습된 단어, 위치에 대한 정보와 학습된 벡터 값을 담고 있다. 후처리 과정에서는 특징들을 바탕으로 학습된 단어들을 최종 연관단어 결과물로 뽑아낸다.

4. 실험 및 평가

실험은 사람이 직접 일주일 분량의 뉴스기사를 읽고 추출한 연관단어와 본 시스템을 통해 생성된 연관단어를 비교하여 성능 평가를 진행하였다. 성능 평가는 평가방법에 널리 이용되는 ROUGE(Recall-Oriented Understudy f

or Gisting Evaluation)[9]를 이용하였다.

본 논문의 시스템과 최근 연관단어추출에 많이 사용되는 LDA를 사용하여 연관단어를 추출하는 시스템을 ROUGE를 통해 정확률, 재현율, F-Measure를 평가하였다. 실험에 쓰인 데이터는 각종 검색엔진에서 제공하는 데이터를 사용했다. 주제어 관련 검색 결과로 나온 뉴스, 블로그 등의 다양한 텍스트 데이터를 수집하여 Word2Vec 학습에 사용했다. 합성곱 신경망의 학습과정에는 주제어 관련 검색일 기준 최근 1주일 분량의 뉴스 검색결과들만 수집하여 사용했다. 자료들을 학습시켜 나온 연관단어들을 바탕으로 LDA와의 성능을 비교한 결과는 다음과 같다.



(그림 4) ROUGE-1 성능평가 그래프

	정확률	재현율	F-Measure
딥러닝 기반 연관단어 추출	0.4278	0.3896	0.4078
LDA	0.3969	0.3615	0.3783

<표 1> ROUGE-1 성능평가 수치표

(그림 4)와 <표 1>에서 볼 수 있듯이, LDA를 사용하여 생성된 연관단어보다 본 논문의 시스템(딥러닝 기반 연관단어 추출)을 사용했을 때, 전체적으로 더 높은 정확률과 재현율 그리고 F-Measure 값을 보이는 것을 확인할 수 있다.

5. 결론

본 논문에서는 Word2Vec를 통해 검색엔진에서 얻은 다양한 텍스트 데이터를 분석하여 벡터 테이블을 생성하였다. 생성된 벡터 테이블을 바탕으로 사용자가 입력한 주제와 관련된 일주일 분량의 뉴스기사들의 단어들을 벡터 테이블을 통해 벡터화 시키고 합성곱 신경망을 통해 특징 벡터를 학습하여 연관단어를 생성하는 시스템을 제안하였다. 실험을 통해 기존 연관단어 연구에서 많이 쓰이던 LDA와 비교를 하여 보다 더 높은 성능을 보임을 확인했다. 이를 통해, 텍스트 문서를 분석하여 내용을 대표할 수 있는 연관단어를 추출성능에 큰 영향을 미친다는 사실을 알 수 있었고, 앞으로의 연구들에도 크게 기여할 수 있을 것이라고 본다. 본 시스템은 영어 기반 텍스트 문서에 적용하여 사용하였다. 현재 딥러닝을 통해 한글 문서를 분석하여 연관단어를 생성하는 연구는 많이 진행되지 않았다. 후속 연구로 한글 문서를 딥러닝을 통해 분석, 연관단어를 생성해 내는 연구를 진행할 예정이다.

참고문헌

[1] Yang, Jinqiu, and Lin Tan. "SWordNet: Inferring semantically related words from software context." Empirical Software Engineering 19.6 (2014): 1856-1886

[2] Das, Pradipto, Rohini K. Srihari, and Jason J. Corso. "Translating related words to videos and back through latent topics." Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013

[3] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." Nature 521.7553 (2015): 436-444

[4] Severyn, Aliaksei, and Alessandro Moschitti. "Learning to rank short text pairs with convolutional deep neural networks." Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2015

[5] Kim, Yoon. "Convolutional Neural Networks for Sentence Classification", Empirical Methods on Natural Language Processing, 2014

[6] Dos Santos, Cícero Nogueira, and Maira Gatti. "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts." COLING. 2014

[7] Goldberg, Yoav, and Omer Levy. "word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method." arXiv preprint arXiv:1402.3722 (2014)

[8] Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines." Proceedings of the 27th international conference on machine learning (ICML-10). 2010

[9] Lin, Chin-Yew, and Franz Josef Och. "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics." Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp.605, 2004