Cohesion Devices 를 이용한 학습 적용 방법과 성능 개선을 위한 실험

김용훈*, 정목동*
*부경대학교 컴퓨터공학과
e-mail : kimyhjava@pukyong.ac.kr, mdchung@pknu.ac.kr

Test on Learning Method for Improving Performance Using Cohesion Devices

Yonghoon Kim*, Mokdong Chung*
*Dept. of Computer Engineering, Pukyong University

요 약

현재의 정보 검색 및 문서를 분류하는 기법에 대하여 신경망을 이용한 정보검색 모델에 대한 연구가 활발히 진행되고 있으며, 간단한 문장에 대한 주제어 분석에서부터 장문에 해당하는 수필 등의 문서를 분류하는 기술이 요구되고 있으며, 이를 실현하기 위한 다양한 알고리즘을 적용하거나, 단어 및 문서에 가중치를 적용하거나, 문서에서의 특이 값을 구하고, 이를 분석하는 방법에 대하여 정보화가 가속화 되면서 정확한 문서에 대한 이해가 요구되고 있다. 이러한 연구와 직접적으로 관련된 단어의 빈도에 대한 논의는 사회과학의 영어학습에 대한 연구 또는 순수 언어에 대한 연구에 머물러 있다. 이에 본 연구에서는 영문에서의 응집장치를 이용하여 문장에서의 중요 단어에 대한 빈도를 합리적으로 증가시켜 문장의 의미를 더 정확하게 분석 할 수 있는 기법에 대하여 제시하고자 하며, 본 논문에서는 영문 수필 사이트의 분류를 추측하고 이를 자동 분류 할 수 있는 방법에 대하여 제시하고자 하며, 이를 구현하여 문서의 의미에 대한 연구에 기여하고자 한다.

1. 서론

데이터 분석에서 분류에 따라 정형데이터와 비정형데이터로 나누며, 그 학습방법에 따라 지도학습과 비지도 학습으로 구분하고 있으며, 현재 비지도 학습을 이용한 비정형데이터 분석은 더욱 활발하게 진행되고 있다. 비정형데이터는 그 의미가 정확하게 구분되지 않은 데이터들을 의미 하며, 이미지, 소리 및 문장이 있다. 특히, 텍스트 관련 비정형데이터 분석 중에서 정보검색(Information Retrieval) [1]은 정보화 시대에 중요한 요소이다. 정보검색은 쿼리에 해당하는 응답으로 검색 결과의 순위를 지정하는 것을 의미한다. 정보검색에서의 지도학습은 문서에 대한 Label 과 문서를 한 쌍으로 저장하고 이를 활용하는 것을 말하며, 비지도 학습은 특정쿼리에 해당하는 정보를 문서의 특징으로 학습하여 분류하는 것을 의미한다. 이와 관련한 전통적 분석 모델로 LSA(Latent Semantic Analysis) [2]가 있으며, 확률적 기법을 적용한 LDA(Latent Dirichlet Allocation) [3]가 있다. 최근 이미지 분석의 CNN(Convolution Neural Network) [4]과 같은 신경망의 관심이 높아짐에 따라 정보검색에 이를 이용한 분석이 높아지고 있다. 정보검색에서 무엇보다 중요한 문제는 문서의 특징을 정확하게 구분하는 것이며, 단어의 유사성과 관련하여 N-gram 및 skip-gram 을 이용한 CBOW(Continuous Bag-of-words) [5]과 주변단어의 연관성을 이용한 word2vec [6] 및 문서 길이를 추가로 적용한 Paragraph2vec [6] 이 있다. 하지만, 단어의 빈도와 문서의 빈도를 기반으로 하는 분석에서 단어의 빈도는 중요한 부분이며, 이에 대한 논의는 아직 부족하다.

본 논문에서는 단어에 대한 빈도와 관련하여 영문학에서 문장 구성 및 작문과 밀접한 응집장치를 바탕으로 문서에서 대체된 단어를 추정 및 삽입하여 문서의 의미를 높이고자 하며, 이를 통하여 더 정확한 문서의 특징을 분석하고 이를 학습하여 이와 비슷한 문서를 예측하는데 기여하고자 하며, 2 장에서 관련연구와 3 장에서 제안하는 시스템구조를 설명하고 4 장에서 시스템에 대한 구현과 평가를 5 장에서 결론과 향후 연구에 대하여 설명하고자 한다.

2. 관련연구

2.1 응집론

구조적인 단위가 두 개 이상 모여서 하나의 텍스트를 구성할 때, 문제가 될 수 있는 단위들

사이의 관계에 대하여 의미상의 일관성이 존재하며, 이 일관성을 언어형식에 의해서 표준화한 것을 응집(Cohesion)[7]이라 하며, 이러한 응집장치들이 어휘와 문법 장치로 작용한다고 하여 어휘적 응집장치와 문법적 응집 장치로 나누고, 어휘적 응집장치는 동일어, 동의어, 상· 하위어 등을 반복하여 사용함으로써 앞의 정보와 논리적으로 연결되어 글의 응집력을 높이는 방법들이 있고, 문법적 응집장치로는 지시어를 사용하여 앞에 나온 정보를 지시하거나, 대용, 생략하여 앞의 정보와 연결되어 있음을 보이거나, 접속시키는 방법[8]을 말하며, 대표적 응집장치 유형 분류는 표 1 과 같으며, 응집장치는 지시(Reference), 대용(Substitution), 생략(Ellipsis), 접속(Conjunction), 어휘(Lexical)로 구분한다 [9].

<표 1> 응집장치 유형 분류

구분	내용
지시	대명사, 지시사, 정관사, 비교사
대용	예) Do you have a pen? Yes, I do.
생략	예)Would you like a cup of coffee? Yes, I would.
접속	부가, 대조, 인과, 시간, 전환
어휘	동일어, 동의어, 상· 하위어

2.2 Tensorflow Softmax

TensorFlow 는 Data flow graph 를 사용하여 수치연산을 하는 open-source software library 이며, Graph 의 Node 는 수치 연산을 나타내고 Edge 는 Node 사이를 이동하는 다차원 데이터 Tensor 를 나타낸다. server 혹은 mobile 디바이스에서 CPU 나 GPU 를 사용하여 연산을 구동시킬 수 있으며, 주로 Machine Learning 과 DNN (Deep Neural Network)연구를 목적으로 Google 의 AI 연구 조직인 Google Brain 팀에 의해 개발되었다 [10].

Softmax 또는 Softmax regression 은 기존 Neural Network 의 출력에 Sigmoid 함수를 사용하지 않고 확률분포로 대체한 것과 같다.

2.3 Neural Network

Neural Network 또는 Artificial Neural Networks 은 분류와 예측을 위한 모형을 말하며, 반응변수의 복잡한 관계를 파악하는 매우 유연한 방법으로 Multilayer Feedforward Networks 가 가장 일반적이며, 전달함수로는 선형함수, 지수함수, Logistic/Sigmoid 함수 등이 있으며, Sigmoid 함수가 대표적으로 쓰인다. 최초 Random 값으로 지정된 가중치를 Back propagation 으로 갱신하여 예측확률과 이 값에 기반한 분류 값을 계산한다 [11].

2.4 Word2vec

벡터 공간에서 단어를 분산 표현하면 유사한 단어를 그룹화하여 자연어 처리 작업에서 더 나은 성능을 얻을 수 있도록 알고리즘을 학습하는 데 도움이 되며, 단어 표현의 가장 초기 사용 중 하나는 1986 년 Rumelhart, Hinton 및 Williams [12]의 연구가 대표적이라 할 수 있다. 이 아이디어는 상당한 부분에서 만족스러운 성공을 거두어 통계 언어 모델링에 적용되어왔다 [13]. 후속 작업에는 자동 음성 인식 및 기계 번역 [14] 응용 프로그램 및 광범위한 NLP 작업이 포함되고 있다.

3. 응집장치 적용

연구에서 응집을 적용함으로써 의미분석에서 정확성을 높일 수 있음을 증명하기 위하여 기존 연구에서는 두 가지로 실험을 진행하였다. 하나는 문서 의미에 대한 특징을 추출하고 이를 학습하여 문서를 구분하는 실험이며, 하나는 문서에서 단어에 대한 의미를 구분하기 위해 사용되는 Word Embedding 이다.

이를 적용하기 위한 데이터는 Wikipedia 에서 사랑(Love) 미움(Hate), 전쟁(War), 평화(Peace), 전쟁과 평화(War and Peace)와 관련된 말뭉치를 사용하였으며, 응집장치가 적용된 데이터와 응집장치가 적용되지 않은 데이터를 구분하여 두개의 데이터 집합을 작성하였으며, 응집장치 중 식별하기 쉬운 지시사 및 대명사를 중심으로 해당하는 단어를 삽입 하였으며, 그림 2 와 같이 지시사 this 에 대하여 이전문장에서 추정할 수 있는 a variety of different emotional and mental states 를 삽입함으로써 응집장치를 적용하였다.

~ An example of **this** range of meanings is ~
~ An example of **this** **(a variety of different emotional and mental states)** range of meanings is ~
~ Love in **its** various forms acts ~
~ Love in **its** **(familial love, friendly love, romantic love, divine love)** various forms acts ~
~ deaths since **its** start, is World War II~
~ deaths since **its**(The deadliest war) start, is World War II ~

(그림 1) 응집장치를 이용한 단어 삽입의 예

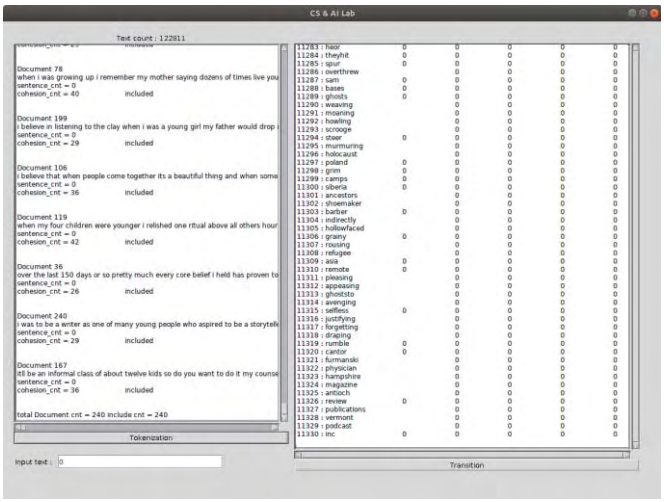
문서구분에서는 문서의 특징을 추출하고 이를 학습하여 비슷한 문서를 추정하는 것으로 말뭉치를 각 대상 단어에 대하여 구분하여 적용하였다. 전체 37 개의 말뭉치 중 27 개의 말뭉치로 학습하고 10 개의 말뭉치를 검증데이터로 사용하였으며, 응집장치를 적용하지 않은 4,085 개 단어와 응집장치를 사용한 4,109 개 단어에 해당하는 1,599 토큰을 적용하였다. 또한 tf-Idf 를 이용한 단어 빈도와 관련하여 일반적으로 사용하는 식은 (1)과 같으며, 기존 실험에서는 $\log\frac{N}{n_t}$ 에서 모든 문서에 단어가 출연하면 0 되는 문제와 관련하여 이 문제점을 해결하기 위하여 조정값 λ를 추가하여 값이 0 이되는 것을 방지하였다. 이것은 LSA 와 같이 완전한 특이 값을 분석하기 위하여 관사와 같은 의미와 관계없는 단어를 제외 시키기 위한 것이지만 본 연구에서는 각 단어의 출현빈도에 중심을 두고 식 (2)와 같이 적용하였다 [15].

$$w_{t,d} = \begin{cases} 1 + \log tf_{t,d} \cdot \log \frac{N}{n_t} & if\ tf_{t,d} > 0 \\ 0 & otherwise \end{cases} \quad (1)$$

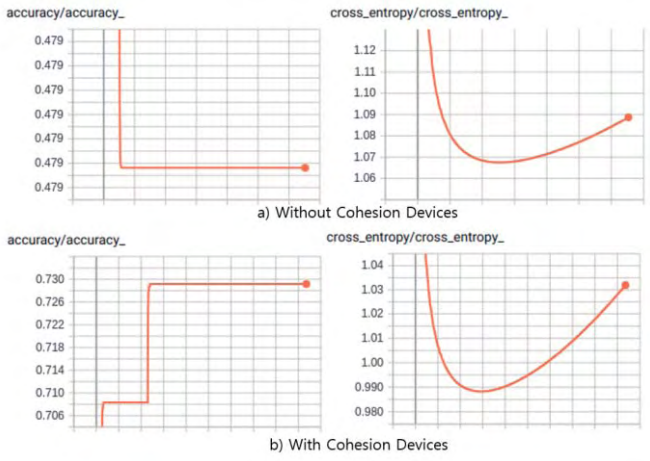
$$w_{t,d} = \begin{cases} 1 + \log tf_{t,d} \cdot \log \frac{N+\lambda_d}{n_t} & if\ tf_{t,d} > 0 \\ 0 & otherwise \end{cases} \quad (2)$$

$$\lambda_d = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}(d_i - E(d))^2\right)}$$

본 연구에서는 this I believe 사이트의 240 개 문서, 122,811 단어, 11,330 토큰을 이용하여 총 4 개의 분류에 속하는 문서를 분류하는 실험을 진행하였으며, 그림 2 는 학습 테이블 생성의 예이다.



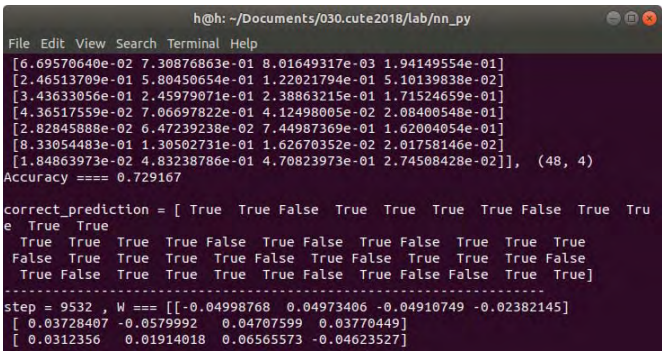
(그림 2) 문서 분류 및 학습테이블 생성의 예



(그림 3) Cohesion Devices 포함과 미포함의 결과

240 개의 문서를 테이블로 단어빈도를 계산하고 이를 학습하였으며, 실제 결과는 기존의 결과보다 좋지는 않았다. 수필의 경우 Wikipedia 에서 추출한 문서보다 같은 분류에 속하는 문서에서의 단어 빈도가 많이 떨어지고 실제 분류하는데 어려움이 있음을 알 수 있었다. 하지만, Cohesion Device 를 사용한 경우와 그렇지 않은 경우의 학습에서는 많은 차이점을 보여 주며, 그림 3 과 같다.

그림 3 에서 a)는 Cohesion Devices 를 적용하지 않은 결과이며, b)는 Cohesion Devices 를 적용한 결과에서의 정확도와 엔트로피를 보인다. 엔트로피가 상승하는 구간까지 학습하였으나, a)의 경우 48%를 보이며, b)의 경우 최대 72.9$의 결과를 확인할 수 있었으며, 최대 정확도까지 9,532 회 학습을 진행하였고 그림 4 와 같다.



(그림 4) 정확도에 대한 실험 결과의 예

6. 평가

실험에서 정확도에 대한 공정성을 위하여 240 개의 문서를 무작위 배열하고 학습데이터와 검증데이터를 8:2 로 구분하여 총 5 회에 걸쳐 학습을 진행하였으며, 결과는 표 1 과 같다.

표 2. 실험에 대한 결과 비교

구분	Wiki	This I believe				
	기존	#1	#2	#3	#4	#5
With Cohesion	60%	64%	66%	73%	65%	63%
Without Cohesion	90%	53%	52%	57%	56%	48%

표 1 에서 Wikipedia 의 결과와 같이 만족 스럽지 못하지만, 평균 70%정도의 정확도를 확인 할 수 있었고, Cohesion Devices 를 사용하지 않은 경우는 평균 54%의 정확도와 비교하였을 때 16%의 오차를 확인 할 수 있었다. 특히 수필과 같이 주제가 같은 경우에도 개인의 특성에 따라 구사하는 단어가 많이 다름으로 인하여 학습을 적용하여도 정확하게 구분할 수 있는 수준은 아니지만, 대명사 및 지시사를 적용하였지만, 적용하지 않은 경우보다 높은 정확도를 확인 할 수 있었다.

7. 결론

본 논문에서는 영문에서의 문장간의 관련 여부를 판단하는 기준이 되는 응집장치를 이용하여 단어에 대한 빈도를 증가시켜 문서에 대한 추론 가능성을 높이고자 하였고, 응집장치를 사용함으로써, DNN 에서의 학습결과는 높아지는 것을 알 수 있었다. 또한, 기존의 단어 빈도 측정방법에서 문서간에 해당하는 단어의 가중치를 추가로 적용함으로써 더

높은 학습효과를 얻을 수 있었으며, 수필과 같은 개인적 특성이 반영된 문서에 대하여 분류 할 수 있음을 일부 확인 할 수 있었다. Cohesion Devices 의 종류가 많지만, 일부분을 적용한 것은 언어학적 구분으로 전제된 다른 장치를 분석부분에 적용하는 어려움이 있을 것으로 보여진다. 하지만, 대명사나 지사사와 같이 구분이 쉬운 단어를 이용하여 학습 정확도를 높일 수 있음을 확인 할 수 있다.

향후 연구에서는 개인의 특성을 고려하여 사이트의 많은 데이터를 적용하여 학습하고자 하며, 총 71 가지의 분류를 적용하기 위하여 점진적으로 추가하는 부분에 대하여 연구하고자 하고, 또한 정확도에 대한 문제점을 개선하고자 한다.

참고문헌

[1] M. Bhaskar, and N. Craswell. "Neural Models for Information Retrieval," arXiv preprint arXiv:1705.01509, 2017.

[2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, “Indexing by latent semantic analysis,” JASIS 41, 6 pp. 391–407, 1990.

[3] L. D. Tyson Who's bashing whom?: trade conflict in high-technology industries, Peterson Institute, 1993.

[4] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on. IEEE, 2014.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781 2013.

[6] X. Rong, “word2vec Parameter Learning Explained,”arXiv preprint arXiv:1411.2738, 2014.

[7] 성창섭, "영어 텍스트의 응집 현상," 새한영어영문학, Vol. 15, pp. 137-157, 1982.

[8] 안수진, "한국인 영어 학습자들의 구어 및 문어담화에 나타난 응집장치 사용 양상," 새한영어영문학, Vol. 59, No. 2, pp. 163-186, 2017.

[9] K. Halliday, and R. Hasan, Cohesion in English, London, Longman, 1976.

[10] Google Inc., https://www.tensorflow.org.

[11] S. Galit, P. C. Bruce, M. L. Stephens, and N. R. Patel, Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner, John Wiley and Sons, Inc., 2016.

[12] David E Rumelhart, Geoffrey E Hintont, and Ronald J Williams. Learning representations by backpropagating errors. Nature, 323(6088):533–536, 1986.

[13] Yoshua Bengio, R´ejean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. The Journal of Machine Learning Research, 3:1137–1155, 2003.

[14] Holger Schwenk. Continuous space language models. Computer Speech and Language, vol. 21, 2007.

[15] Y. Kim, H. Hong and M. Chung, “Application of Cohesion Devices for Improvement of Distributional Representation,” proceeding of The 14th International conference on Multimedia Information Technology and Applications(MITA2018), June 28-30, 2018, Shanghai University of Engineering Science, China, pp. 84-87, 2018.