|  |
| --- |
| **딥러닝 기반 한국어/영어 감성 및 주제 분석**  유근영, 김동혁, 김세윤, 이호현  고려대학교,디지털금융융합과정  keunyoung.yu@gmail.com; swkdh@nate.com; crazytom08@naver.com; hochi575757@gmail.com  **Deep Learning Based Korean/English Sentiment Analysis**  Keunyoung Yu, Donghyuk Kim, Seyun Kim, Hohyeon Lee  Korea University, Digital Finance Convergence Engineering  **요 약**  자연어로 표현된 문장 속 문맥을 분석한다는 것은 주체인 사람의 주관을 파악한다는 것이기에 단순한 통계적 빈도 분석으로는 한계에 봉착하기 마련입니다. 더욱이 자연어를 설계한 인간이 규칙으로 모델을 구분해 두더라도 수많은 예외 사항을 별도로 처리하기란 불가능한 일입니다.  하지만 분산분석을 통해 인간이 자연어에 무의식적으로 부여하고 있는 특징들에 한걸음 더 나아갈 수 있게 되었고, 자연어의 분포가설이라는 큰 가정은 베이지안 확률을 통해 단어에서 문맥으로 나아가는 게기가 되었습니다.  본 논문 역시 위에 열거한 자연어처리의 핵심 아이디어를 공유하고 있는 가운데 발전된 딥러닝 학습법을 적용해 자연어 감성 분석 성능을 높임을 목적으로 합니다.  주제어: 문맥 분석, 감성 분석, 주제 분석, LSTM |

**1. 서론**

자연어 모델(Natural Language Model)이란 단어 혹은 문장에 어떤 판단을 위한 확률을 할당하는 모델을 말합니다. 그 판단은 감성, 주제, 다른 언어 등 다양합니다. 본 논문에서는 감성이라는 판단에 부합하는 확률을 할당하는 자연어 모델을 수립함에 목적이 있습니다.

자연어 모델의 종류는 큰 틀에서 빈도기반, 예측기반, 빈도와 예측 결합기반으로 나뉘게 됩니다.

빈도기반 모델은 코퍼스(Corpus)-표본집단으로 자연어(Natural Language)-모집단을 근사하는 것에 목표를 둔 방법입니다. 대표적인 모델로는 CBOW, Skip-gram, TF-IDF가 있습니다.

반면, 예측기반 모델은 단어를 벡터로 표현함으로 빈도기반 모델의 희소 문제를 해결하고 있습니다. 단어를 벡터로 표현하는 방법을 분산표현[2]이라고 하며, 분포가설[1] 가정 하에 만들어진 표현입니다. 비슷한 위치에서 등장하는 단어들은 비슷한 의미를 가진다는 것입니다. 이러한 측면에서 예측기반 모델과 빈도기반 모델이 공유하는 가설은 같음을 알 수 있습니다. 다만, 빈도기반 모델의 경우 비슷한 의미를 가지는 단어들을 연결하는 방법이 '빈도'라는 특징에 제한 되었으나, 예측기반 모델의 경우 단어를 벡터로 표현함에 따라 인간이 자연어에 무의식적으로 부여하고 있는 특징들을 포함하고자 하는 방식입니다. 대표적인 모델로는 Word2Vec과 FastText이 있습니다.

최근에는 빈도기반 모델과 예측기반 모델을 결합한 방식 역시 많이 활용되고 있습니다. 대표적으로 Glove(Global Vectors for word representation), ELMo (Embedding from Language Model)가 있습니다.

본 논문의 경우에는 단어, 문장 표현이 아닌 단어, 문장이 이루어진 문맥 속 감성을 분석함에 목적이 있기에 기존에 학습된 단어벡터를 활용하기 보다는 최적화 모델에 초점을 맞추어 LSTM(Long Short-Term Memory)을 활용하였습니다.

연구의 방향을 NSMC(Naver Sentiment Movie Corpus) 데이터[3]와 Friends 대본 데이터[4]에 LSTM 모델을 적용하여 첫째, 문맥을 반영한 독자적 단어 벡터 스페이스를 구축하는 것, 둘째, 코퍼스(Corpus)의 문맥정보를 소실없이 학습하는 것, 마지막, 문맥 판단의 목적(감성, 주제)에 따라 모델을 설계하는 것으로 정하였습니다.

본 논문의 연구를 통해 LSTM을 통한 학습 방법이 한글과 영어 두 언어에 적용하여 학습률을 높일 수 있음을 보이고 있습니다.

**2. 제안하는 방법**

**2.1 Sequence Model**

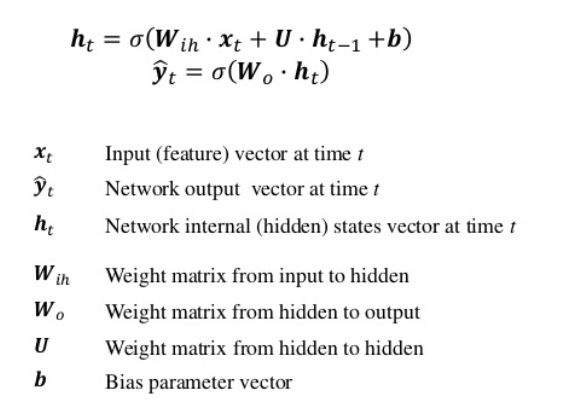
문맥과 목적(감성/주제)간의 패턴을 학습하는 모델을 적용함에 있어 가장 우선하여 고려한 점은 문장 역시 순서가 있는 시퀀스 데이터라는 것입니다.

시간축 인덱스를 기준으로 정리된 데이터만이 시퀀스가 아니라 순서가 존재하는 데이터는 모두 시퀀스 데이터로 간주해야 하기에 시퀀스를 처리할 수 있는 모델을 적용해야 합니다.

따라서 입력받은 데이터들의 순서까지 고려하는 시퀀스 모델(RNN, LSTM, GRU 등)이 우선 고려대상이 되었습니다.

시퀀스 처리 모델은 인공신경망 내 은닉계층의 유닛들이 방향성 사이클(Directed Cycle)을 포함하는 구조를 가지기 때문입니다.

이러한 구조는 인공신경망으로 하여금 은닉계층에 저장되어 있는 과거의 정보들을 현재의 입력값과 결합하여 사용할 수 있게 함으로써 입력값 시퀀스에 대해 시퀀스 내부 유닛간의 상호적인 정보를 어느 정도 유지 시킨 은닉변수 시퀀스로 비선형 변환을 시킬 수 있습니다.



[그림1] RNN 모델 기본 구조

**2.2 LSTM**

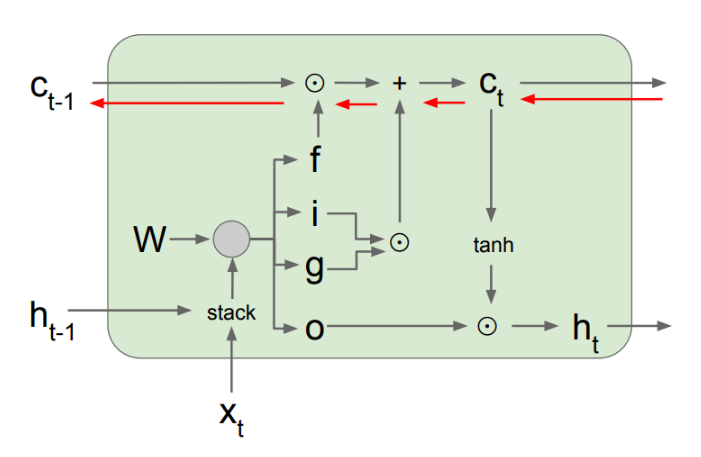
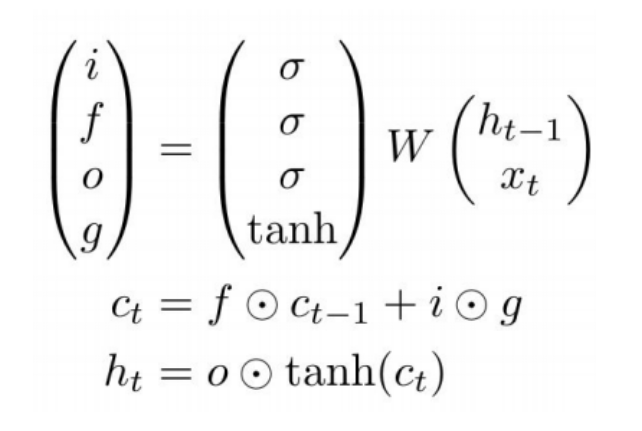
해당 논문에서 특별히 LSTM 모델을 적용한 것은 시퀀스 모델이 가지는 단점인 장기 의존성 문제(the problem of Long-Term Dependencies)를 해결하기 위함에 있습니다.

장기 의존성 문제는 문장의 길이가 길어질 때, 최적해를 찾아가는 역전파(Back Propagation) 알고리즘의 연쇄법칙(Chain Rule) 계산 과정에서 Sigmoid, tanh와 같은 활성화 함수를 여러번 거침에 따라 기울기(Gradient)가 극도로 작아지는 문제를 의미합니다. 즉, 문장이 길어질 경우 중요 정보가 소실되는 문제가 발생하는 것입니다.

LSTM은 은닉층의 메모리 셀에 입력 게이트, 망각 게이트, 출력 게이트를 추가하여 불필요한 기억을 지우고, 기억해야할 것들을 정하면서 장기 의존성 문제을 해결하고 시퀀스 정보를 저장하는데 탁월한 성능을 보입니다.

한편, LSTM은 매우 다양한 유형이 존재합니다. 단방향(unidirectional)/양방향(bidirectional), 단층구조(multi-layered LSTM)/다층구조(stacked LSTM), Many-to-One/Many-to-Many 등이 있습니다.

본 연구에서는 학습문장과 학습목적에 따라 영화평 감성 분석 데이터인 NSMC(Naver Sentiment Movie Corpus)[3]를 해결하기 위하여 LSTM을 Many to One 구조로 설계하였고, Friends 대본 데이터[4]는 Many to Many 구조로 설계하였습니다.



[그림2] LSTM 모델 기본 구조

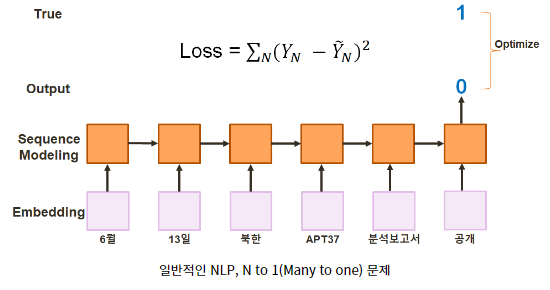
**2.2 Many to One**

문장 벡터의 N Features를 받고 Binary 분류 Class를 출력으로 도출하는 경우에 해당하는 구조입니다.

따라서 모델의 최적화 목적함수로도 Binary Cross Entropy를 주로 사용합니다.

대표적인 Many to One 문맥 분석의 예로는 추천/비추천, 긍정/부정 등이 있습니다.

해당 연구에서는 영화평 감성 분석 데이터 NSMC(Naver Sentiment Movie Corpus)[3] 문맥 분석을 Many to One 구조를 통하여 진행하였습니다

[그림3] Many to One 구조

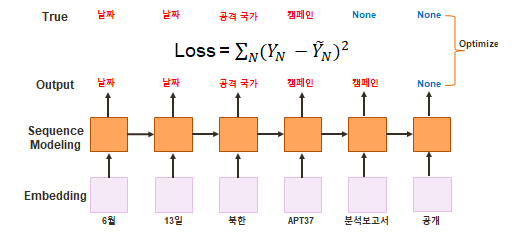
**2.2 Many to Many**

문장 벡터의 N Features를 받고 Multi 분류 Class를 출력으로 도출하는 경우에 해당하는 구조입니다.

따라서 모델의 최적화 함수로는 Categorical Cross Entropy를 사용하고 있습니다.

대표적인 Many to Many 문맥 분석 예로는 뉴스 주제 카테고리 분류, 영화 평점(1~5) 분류, 다양한 감정(기쁨, 슬픔, 화남, 분노 등) 분류가 있습니다.

해당 연구에서는 Friends 대본 데이터[4] 감성 분석 문제를 Many to Many 구조로 진행하였습니다.

[그림4] Many to Many 구조

**3. 실험**

**3.1 실험 세팅**

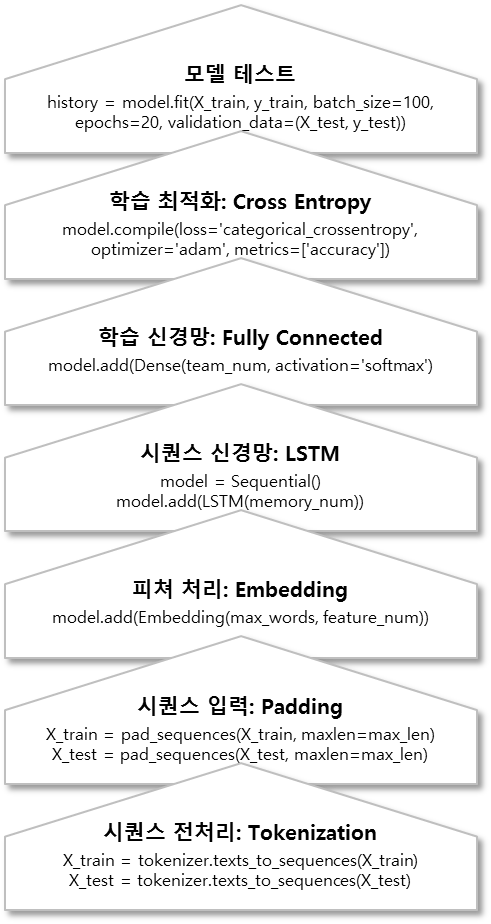
실험에서는 사용된 데이터는 영화평 감성 분석 데이터인 NSMC(Naver Sentiment Movie Corpus)[3]과 Friends 대본 데이터[4] 입니다.

[그림5] 실험 세팅

본 논문에서 제시한 LSTM 모델 및 자연어 처리와 관련된 모델은 Python 언어 및 텐서플로우 케라스 라이브러리로 작성되었습니다. 해당 실험은 Google Colab 환경에서 전처리부터 최적화까지 진행되었습니다.

**3.2 실험 과정**

실험의 전반적인 프로세스는 다음과 같이 진행하였습니다. 1. 데이터 이해 2. 문장 전처리 3. 문장 토큰화 4. 문장 임베딩 5. 문맥 신경망 6. 학습 신경망 7. 학습 최적화 8. 모델 테스트

[그림6] 실험 과정

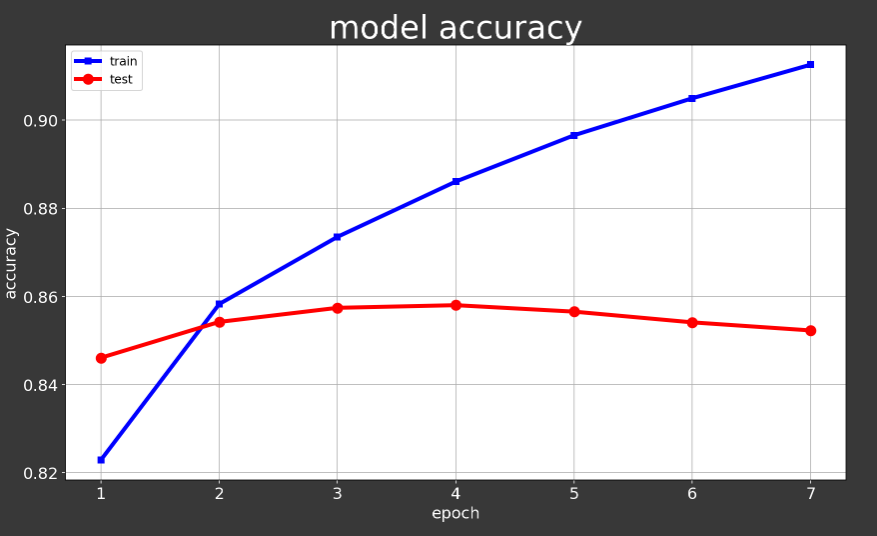
**3.3 실험 결과**

해당 문맥 분석 실험은 큰 틀에서 문장 임베딩, 문맥 학습, 최적화 모델 세 가지에서 성능을 좌우하게됩니다.

문장 임베딩의 경우 패딩을 통해서 300단어로 맞추어서 학습을 진행하였고, 피쳐는 120개로 설정하였습니다.

문맥정보를 파악하는 LSTM 신경망 층은 과적합을 막기 위해 2개의 층만 추가로 구축하였습니다.

최적화 모델을 쌓는 신경망 층은 RMSProp 모델을 사용하였습니다. 해당 모델은 Gradient Descent 기반 최적화 모델 가운데 Per-parameter adaptive learning rate 방법으로서 각 파라미터에 대해 adaptive한 학습 속도를 적용하여 효율적인 방법론으로 알려져 있습니다.



[그림7] 네이버 영화 감성 분석 모델 테스트 결과

[그림8] 네이버 영화 감성 분석 모델 테스트 결과

**4. 결론**

본 논문에서는 문맥 분석 모델 중 하나인 LSTM에 대하여 소개하고 한국어와 영어 데이터를 이용하여 문맥 분석을 진행하였습니다.

먼저, 영화평 감성 분석 데이터인 NSMC(Naver Sentiment Movie Corpus)[3]과 Friends 대본 데이터[4] 감성 분석 문제를 해결하기 위하여 LSTM에 기반한 딥러닝 모델을 제안하였습니다.

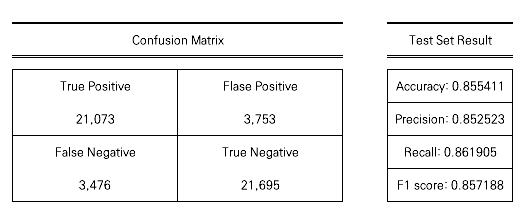
실험 결과, 본 논문에서 제안한 방법은 Accuracy: 85%, Precision: 85%, Recall: 86%, F1 score: 0.85 기록하였습니다.

향후 연구로는 BERT를 적용한 문맥 분석 모델을 구축할 예정이며, 단일 언어 내 문맥 분석 뿐만 아니라 이종언어간의 문맥 분석에도 적용해 볼 예정입니다.

**감사의 글**

이 논문은 2020년도 고려대학교 디지털금융융합과정 인공지능과 자연어처리 임희석 교수님과 연구실 튜터분들의 지원을 받아 수행된 연구입니다.

**참고문헌**

1. Distributional structure, Zelling Harris, 1954
2. Learning distributed representations of concepts, Geoffrey E. Hinton, 1986
3. https://github.com/e9t/nsmc
4. http://doraemon.iis.sinica.edu.tw/emotionlines/download.html
5. http://colah.github.io/posts/2015-08-Understanding-LSTMs/
6. LSTM 언어모델 기반 한국어 문장 생성, The Journal of Korean Institute of Communications and Information Sciences '16-05 Vol.41 No.05, 김양훈, 황용근, 강태관, 정교민, 2016