LSTM 모델을 통한 국문 기사 감성 분류 시스템

The Sentiment Classification of News Articles using LSTM

저자 (Authors)
유은조, 이지현, 박소영
Eun-joe You, Ji-hyeon Lee, So-young Park

출처 (Source)
한국정보과학회 학술발표논문집 , 2018.12, 1949-1951(3 pages)

발행처 (Publisher)
한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07614149

APA Style
유은조, 이지현, 박소영 (2018). LSTM 모델을 통한 국문 기사 감성 분류 시스템. 한국정보과학회 학술발표논문집, 1949-1951

LSTM 모델을 통한 국문 기사 감성 분류 시스템

유은조O이지현O박소영
건국대학교 소프트웨어학과
yej330@naver.comOyuio0312@naver.comOsoyoungpark@konkuk.ac.kr

The Sentiment Classification of News Articles using LSTM

Eun-joe YouO Ji-hyeon LeeOSo-young Park
Department of Software, Konkuk University

요 약

정보화 사회에 살아가는 우리는 넘쳐나는 정보를 제공받고 있지만 그 정보를 제공하는 뉴스나 칼럼의 성향이 극단적으로 갈림으로써 편향된 의견밖에 받지 못하는 경향이 있다. 이러한 불편함을 해소하고 칼럼의 경향을 한 눈에 알아볼 수 있도록 하고자 LSTM을 이용한 뉴스 기사 감정 분류 모델을 제안하였다. 본 연구에서는 수집한 기사 데이터들을 형태소 분리하여 단어 사전을 만들고, 우호적인 기사는 1, 비판적인 기사는 0의 값으로 라벨링하여 해당 모델에게 긍정적, 부정적 성향을 학습시켰다. 이후 사람이 평가한 기사의 우호도와 제안한 모델의 예측 결과를 비교하여 정확도를 평가하였다.

1. 서 론

정보화 사회에 살아가는 우리는 매일 방대한 양의 정보를 접하게 된다. 하지만 그 정보를 전달하는 매체인 뉴스나 칼럼의 성향에 따라 편향적으로 치우친 정보만을 제공받을 위험성이 있다. 이러한 문제를 해결하기 위해 우리는 사용자로 하여금 기사의 객관성을 알 수 있도록 기사가 해당 주제에 대해 우호적인지 비판적인지를 자동으로 분석하는 서비스를 개발하였다.

본 논문에서는 국문으로 이루어진 장문의 기사가 주어지면 LSTM 모델을 이용하여 문서의 감성을 판별하고 긍정적 또는 부정적 경향을 백분율로 제공하는 방법을 제안한다. 또한, 사용자의 편의를 위해 긴 문서의 간단한 요약도 함께 제공한다.

2. 관련 연구

2.1 국문 감성분류 관련 연구

최근 신경망이 발전한 형태인 딥러닝을 이용한 감성분석 연구가 활발하게 이루어지고 있다. 이재준(2018)의 연구[1]에서는 RNN을 이용해 영화 리뷰를 형태소·어절 단위 모형으로 나누어 국문 데이터 감성분석을 실행했다. 본 연구에서는 비교적 긍·부정적 경향이 뚜렷하게 갈리는 영화 리뷰가 아닌 디지털 신문 기사를 대상으로 감성 분석을 실시한다. 또한 사용자가 주제어 또는 키워드를 입력하면, 해당 주제에 대한 신문 기사를 검색한 후, 검색된 기사의 감성을 분석하여 긍정적 또는 부정적 성향 및 세줄 요약을 제시한다.

기사나 칼럼과 같은 여러 문장으로 이루어진 긴 문서는 문장 간 전후 연결 구조가 고려되어야 하기 때문에, 이러한 데이터의 형태를 처리하기에는 RNN[2]이 가장 적합하다. RNN은 다수의 데이터가 순서대로 입력되었을 경우, 기준 시점(t)과 다음 시점(t+1)의 데이터를 연결하여 구성한 인공 신경망(ANN)이다. 그리고 기준 시점(t)에서 기억된 데이터의 중요도를 판단하여 별도의 가중치를 부여한 후 다음 시점(t+1)으로 전달된다. 그러나 이러한 RNN 기법은 기준 시점(t)으로부터 오래된 데이터의 가중치가 소실되어 장기 의존성에 문제가 발생한다. 본 연구에서는 길이가 긴 문서들을 처리해야 하므로, 이러한 장기 의존성 문제가 개선된 LSTM(Long Short-Term Memory) 모델을 최종 학습 모델로 사용한다.

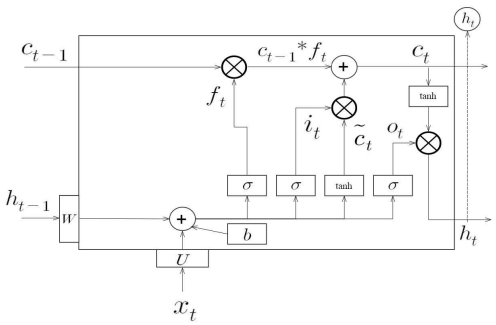

그림 1. LSTM 블록 구조 [2]

LSTM[2]의 핵심은 그림 1에서의 에 해당하는 소자변수(Cell state)에 있다. LSTM은 Long term 메모리와 Short term 메모리로 구성되어 있으며, 소자변수가 두 메모리를 모두 참조하여 Long term 메모리의 데이터를 지속적으로 갱신한다. 따라서 RNN에서 입력 위치 차이가 큰 데이터 사이의 연관성이 소실되는 문제가 발생하지 않는다.

2.2 국문 요약 관련 연구

본 논문에서는 사용자의 편의를 위해 세 줄 요약 기능도 함께 제공한다. 가장 대중적인 요약 알고리즘으로 TextRank[3]가 있다. TextRank는 구글에서 사이트 검색 성능을 높이기 위해 개발한 PageRank[4] 알고리즘을 기반으로 한다. PageRank는 각 사이트를 노드로 보고, 사이트의 호출 빈도를 가중치로 하여 중요한 사이트를 찾아내는 알고리즘이다[4].

TextRank는 노드를 사이트가 아닌 문장으로 대체하고, 문장들 간 명사의 유사도를 가중치로 하여 해당 문서에서 가장 많이 쓰인 명사들을 가진 문장 순으로 중요도를 산출한다. 그러나 이는 요약된 문장들 간의 유사도가 고려되지 않고, 여러 주제를 다루는 긴 문서의 경우 한 가지 주제에 대해서만 요약이 될 수 있다는 문제점이 있다[5]. 이를 보완하기 위해서 LexRank[5]는 문장간 유사도를 분석하여 유사한 문장의 중복을 제거한다. 본 연구에서는 LexRank를 활용하여 국문 기사의 세 줄 요약을 제공한다.

3. 국문 기사 감성 분류 · 요약 시스템

3.1 시스템 모델

제안하는 시스템의 전체적인 동작 기법은 다음 그림 2와 같다.

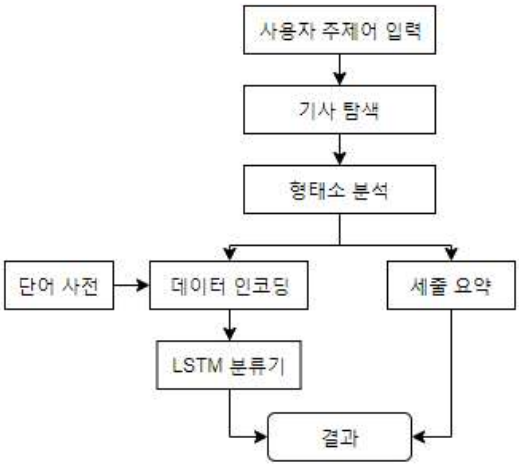

그림 2. 시스템 순서도

시스템은 가장 먼저 사용자에게 주제어를 입력받은 후 해당 주제어와 관련된 기사를 탐색하여 리스트를 출력한다. 사용자가 원하는 기사를 선택하면 시스템 내에서 해당 기사 전체의 형태소를 분석한 뒤 LSTM을 이용한 감성 분류 과정을 거쳐 감성 분석 결과 및 세 줄 요약을 제공한다.

3.2 LSTM 기반 감성 분류기

3.2.1. 형태소 분석

LSTM 감성 분류기 구축을 위해서는 국문 기사의 감성에 대한 학습이 필요하다. 이를 위해서 다양한 주제의 국문 기사를 수집하여 학습 데이터를 생성하였다. 학습 데이터 생성을 위해서, 먼저 국문 기사를 형태소 단위로 분리한 다음, 분리된 형태소 단위에 정수 값을 매핑하는 데이터 인코딩 과정을 수행하였다.

영문 텍스트는 문법별 단어 변형이 크지 않아 띄어쓰기 단위로 구분하기 용이하다. 반면 국문 텍스트는 수많은 어휘 변형과 이런 어휘들의 다양한 조합으로 이루어져 있어 데이터 단위 선정 단계가 필요하다. 그러므로 본 연구에서는 국문에서 가장 작은 말의 단위인 형태소 단위로 구분한다.

3.2.2. 단어 사전 생성

데이터 인코딩을 위해 단어사전을 구축하였다. 본 논문에서는 500개의 신문 기사를 수집하여 모두 형태소 분석을 한 후 빈도수에 따라 내림차순으로 정렬하여 형태소 별 index를 할당하였다. 표 1은 제안한 단어사전에서 최대빈도수를 갖는 14개의 형태소를 보여준다.

표 1. 단어사전

index	형태소	index	형태소	index	형태소
1	하	6	다	11	대하
2	이	7	있	12	북한
3	ㄴ	8	에서	13	ㄴ다
4	는	9	것	14	선언
5	을	10	으로	...	...

3.2.3. 데이터 인코딩

국문 기사가 주어지면, 형태소 분석을 통해 행태소 단위로 분리한 후, 표 1의 단어사전을 이용하여 데이터 인코딩을 수행한다. 표 2는 데이터 인코딩의 예를 보여준다.

표 2. 제안한 단어사전을 기반으로 한 데이터 인코딩 예

2일 국회 경제분야 대정부 질의 김동연 경제부총리 겸 기획재정부 ...	1199, 52, 2599, 431, 52, 56, 33, 46, 26, 105, 72, 114, 25, 867, 170, ...

3.2.4. LSTM 감성 분류기

감성분석을 위해 각 기사마다 0 혹은 1의 분류 값을 출력한다. 라벨 0은 해당 부정적인 성향을 나타내고, 라벨 1은 긍정적인 성향을 나타낸다. 제안한 분류기의 정확도는 4장에서 자세히 분석한다.

3.3 세 줄 요약

본 연구에서는 LexRank 알고리즘을 기반으로 세 줄 요약을 제공한다. LexRank는 비슷한 주제의 문장들은 가까이 있을 확률이 높다는 것에 기반하여 문장 간 유사도를 다음 수식 1에 따라서 결정한다.

$$\text{sim}_{\alpha}(s_i, s_j) = \left(\frac{\max W - |i-j|, 0}{W}\right)^{\alpha} \cos\theta \quad (1)$$

W는 문장 개수 윈도우, α는 감쇠율, s_i는 i번째 문장을 나타내고,, s_j는 j번째 문장을 나타낸다. 이는 코사인 유사도를 이용해 계산된다.

또한 LexRank는 포함되어있는 단어의 중요성에 따라 단어와 문서의 연관성을 계산하는 TF-IDF(Term Frequency - Inverse Document Frequency)벡터[6]를 기반으로 문장들을 클러스터링을 한다. 그리고 각 클러스터 내에서 유사한 문장들의 중복 제거 작업을 하여 여러 주제를 다루는 문서에서 언급이 적었던 주제가 소실되지 않도록 요약이 가능하다.

4. 실 험

4.1. 감성 분류기 정확도 분석

제안한 분류기의 정확도 분석을 위해서 찬반 논란의 여지가 있다고 판단한 다양한 국문 기사를 활용하였다. 특히, 최저임금, 난민수용, 트럼프 미 대통령, 비핵화, JSA 무장해제 등 최신 이슈 18개와 관련된 500개의 기사를 수집하여 LSTM 모델을 학

습시켰다. 각 주제와 주제별 훈련데이터 및 테스트 데이서 개수는 다음 표 3과 같다.

표 3. 주제별 훈련 데이터 및 테스트 데이터의 개수

기사 주제	훈련 데이터	테스트 데이터
최저임금	46개	10개
난민 수용	25개	10개
트럼프 미 대통령	47개	10개
비핵화	57개	10개
JSA 무장해제	27개	10개
기타	314개	30개
총합	516개	80개

각 주제별로 선별된 테스트 데이터들을 모두 모아서 최종 테스트 데이터 세트(FT_SET)를 구성하였다. 먼저 주제 별로 기사를 나누어 모델을 학습시킨 뒤, 각 학습에 사용된 훈련 데이터와 FT_SET에 대한 정확도를 분석하였다. 또한, 516개의 모든 주제에 대한 훈련 데이터로 분류기를 학습시킨 후, 훈련 데이터 및 FT_SET에 대한 정확도도 분석하였으며, 그 결과는 표 4와 같다. 주제 별로 분류기를 학습 시켰을 경우, 해당 주제에 대한 훈련 데이터의 정확도는 모두 80%이상의 정확도를 나타낸 반면, FT_SET에 대해서는 60% ~ 76%의 정확도를 나타내었다. 이는 FT_SET에 훈련에 사용되지 않은 다른 주제에 대한 기사가 포함되어 있으므로, 정확도가 다소 낮게 나온 것으로 판단된다. 반면, 모든 주제에 대한 훈련 데이터로 분류기를 학습 시켰을 경우에는 훈련 데이터에 대한 정확도가 96%로 개선되었고, FT_SET에 대한 정확도도 81%로 개선되었다. 결과적으로 제안한 모델이 81% 이상의 정확도로 기사의 성향을 분류하였다.

표 4. 훈련 데이터 및 테스트 데이터의 정확도 비교

분류기 학습에 사용된 주제	훈련 데이터 정확도	FT_SET 정확도
최저임금	80%	70%
난민 수용	84%	63%
트럼프 미 대통령	85%	68%
비핵화	88%	71%
JSA 무장해제	81%	65%
기타	95%	76%
전체 주제	**96%**	**81%**

또한, 전체 기사로 모델을 학습 시킨 후 주제 별 훈련 데이터 및 테스트 데이터에 대한 정확도 분석을 실시했다. 결과는 다음 표 5와 같다.

4.2. LexRank를 통한 세 줄 요약

LexRank를 통한 세 줄 요약을 수행하였고, 다음 표 6과 같이 문서의 맥락과 주요 주제를 잘 반영한 결과를 도출하였다.

표 6. LexRank 원문과 요약 결과의 비교

원문
미국은 당분간 소련과 긴밀한 관계를 기대할 수 없다. 미국은 소련을 파트너가 아닌 경쟁자로 보아야 한다. 소련은 평화와 안정을 존중하지 않을 것이다. 또 사회주의와 자본주의가 영구히 행복하게 공존할 수 있을 거라고 믿지 않는다. 소련은 모든 경쟁국가를 혼란에 빠뜨리고 그 영향력을 약화시키기 위해 신중하고 집요하게 노력할 것이다. 다행히 소련은 아직 서방에 비해 약한 위치에 있다. 다행히 소련사회는 자신의 모든 잠재력을 약화시킬 결점들을 갖고 있다. 그러므로 미국은 소련이 평화롭고 안정된 세계의 이익을 침해할 기미를 보이는 모든 지역에서 자신감을 갖고 불굴의 군사력으로 소련에 대결하는 확고부동한 봉쇄정책을 추진할 수 있다.
요약
소련은 모든 경쟁국가를 혼란에 빠뜨리고 그 영향력을 약화시키기 위해 신중하고 집요하게 노력할 것이다. 다행히 소련사회는 자신의 모든 잠재력을 약화시킬 결점들을 갖고 있다. 그러므로 미국은 소련이 평화롭고 안정된 세계의 이익을 침해할 기미를 보이는 모든 지역에서 자신감을 갖고 불굴의 군사력으로 소련에 대결하는 확고부동한 봉쇄정책을 추진할 수 있다.

5. 결 론

본 논문에서는 LSTM 모델을 이용하여 국문 기사의 감성을 분석하는 분류기 및 LexRank를 이용한 세 줄 요약 기능을 제안하였다. 국문의 특성을 고려하여, 형태소 분석을 통해 최소 단위로 분리하였고, 제안한 단어 사전을 바탕으로 데이터 인코딩을 진행하였다. 제안한 모델은 18가지의 다양한 주제에 대한 신문 기사에 대해서 81%의 정확도로 기사의 감성을 분류하였다. 향후, 사용자 기사 검색에 따른 실시간 감성 분류 처리를 위해서 보다 효율적인 데이터 인코딩 기법 및 정확도 개선을 위한 학습 모델 강화 기법에 대한 연구를 지속적으로 진행할 예정이다.

6. Acknowledgement

본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 서울어코드 활성화 지원사업 (IITP-2018-2012-1-00593)의 연구결과로 수행되었음. 본 연구는 교육부의 지원을 받아 한국연구재단의 이공분야 기초연구사업의 연구 결과로 수행되었음. (NRF-2014R1A1A3053491).

참 고 문 헌

[1] 이재준, “RNN을 이용한 한국어 감성분석 : 온라인 영화 후기를 중심으로”, 국민대학교 일반대학원 석사 학위논문, 64쪽, 2018.

[2] 김양훈, 황용근, 강태관, 정교민, “LSTM 언어모델 기반 한국어 문장 생성”, 한국통신학회논문지 제41권 제5호, p592-601(10쪽), 2016.

[3] 홍진표, 차정원, “TextRank 알고리즘을 이용한 한국어 중요 문장 추출” 한국정보과학회 2009 한국컴퓨터종합학술대회 논문집 제36권 제1호, p311-314(4쪽),

[4] Brin Sergey, Page Lawrence, “The anatomy of a large-scale hypertextual web search engine”, Computer Networks 제56권 제18호, p3825-3833(9쪽), 2010.

[5] 설진석, 이상구, “lexrankr : LexRank기반 한국어 다중 문서 요약, ” 2016년 동계학술발표 논문집, p458-460(3쪽), 2016

[6]Erkan G 외 1명, “LexRank:Graph-based Lexical Cnterality in Text Summarization”, J.Artif.Intel, vol22, no.1, p457-479(23쪽), 2004.