GRU 알고리즘을 활용한 트위터 뉴스 토픽 추출 알고리즘

윤원철, 남상윤, 박동진, 서재오, 김재성, 이주영, 김재수*
*경북대학교
{yoon1fe, nsun9505, pdj1649, oknan27, jskim94, jasonecc, *kjs}@knu.ac.kr

Twitter news topic extraction algorithm using GRU algorithm

Woncheol Yoon, Sangyoon Nam, Dongjin Park, Jaeoh Seo, Jaeseong Kim, Joo-Young Lee, Jaesoo Kim*
*Kyungpook Natl Univ.

요 약

현재 수많은 대용량 뉴스 빅데이터가 존재하는 상황에서 이를 활용하기 위한 시스템이 없는 상황이다. 대용량 뉴스 빅데이터를 활용하기 위해 문서에서 토픽을 추출하는 과정이 필수적이다. 기존 토픽 추출에서 사용하는 LDA 등의 방식이 사용되고 있지만, 본 논문에서는 주어진 뉴스 기사 데이터 셋에서 Khaii 형태소 분석기를 사용하여 데이터 전처리를 하고 GRU 알고리즘을 사용하여 토픽을 추출하는 모델을 구축하였다.

Ⅰ. 서 론

현재 수많은 언론사에서 매일 수 만개의 대용량 뉴스 빅데이터가 쏟아져 나오고 있다. 하지만 쏟아져 나오고 있는 대용량 뉴스 빅데이터의 핵심적인 단어 혹은 주제를 추출해내는 시스템이 부족한 실정이다.

기존의 토픽 추출에 있어 사용하는 알고리즘은 LDA와 단어 군집화 방법을 이용한 토픽 추출방법이 있다[1]. 하지만 이러한 방법은 특정 토픽이 여러 개의 토픽으로 추출되는 중복 토픽 문제가 발생한다. 또한 추출된 하나의 토픽 내에 여러 토픽이 혼재하는 문제가 발생할 수 있다[2][3].

본 논문에서는 위와 같은 문제점을 해결하기 위해 GRU 알고리즘을 활용하여 주어진 뉴스기사 데이터 셋에서 토픽을 추출하는 알고리즘을 제안한다.

Ⅱ. 본론

A. 트위터 데이터 셋

본 논문에는 모델을 학습시키기 위해 트위터 상에 있는 데이터셋을 사용하였다. 트위터 데이터셋은 총 약 14만개다. 이들 중 트위터 본문에 지도학습 시 타겟으로 사용할 데이터가 한 개 이상 포함되어 있는 데이터셋을 추출한 결과 [표 1]과 같이 약 8만개의 데이터셋을 구축하였다.

row	14554
총 해시태그 개수	486743
추출된 해시태그 개수	127625
row 한 개당 해시태그 개수 평균	3.34407162
추출된 해시태그 비율	26.2202%
추출된 해시태그를 갖고있는 row 개수	87107
추출된 해시태그를 갖고있는 row의 평균 해시태그 개수	1.46515205

[표 1] 트위터 데이터 셋

B. 데이터 전처리

텍스트 분석에 있어서, 데이터 전처리 과정은 매우 중요하다. 전처리 과정에서 부사, 조사 등을 제외한 단어를 추출해내기 위해 형태소 분석이 필수적이다[4]. 본 연구에서 사용하는 데이터 셋들은 띄어쓰기가 없는 중의적 문장이 많다. 따라서 띄어쓰기가 없는 문장을 형태소 분석하는데 높은 적합률을 보인 카카오의 Khaiii 형태소 분석기를 사용하였다[5].

[그림 2] Khaill 형태소 분석 결과

C. GRU 알고리즘

많은 연구에 사용되는 RNN 알고리즘의 경우 이전 스텝의 히든만을 받는 구조이다. 그렇기에 장기 기억을 하기 어려운 단점이 있다. 이런 문제가 발생하는 이유는 RNN Cell을 거치면서 특정 연산을 통해 데이터가 변환되어 일부 정보가 다음 스텝마다 사라지기 때문이다[7]. GRU는 LSTM[8]이 가지는 장점인 장기적으로 기억할 스텝의 정보를 따로 캐싱하는 것을 유지하면서 계산복잡성을 크게 낮춘 기계학습 알고리즘이다 GRU에서는 LSTM에서처럼 Vanishing Gradient Problem[9] 을 해결하면서도 계산에 필요한 Gate의 개수를 줄여 단순화시킨 것이 특징이다[10]. 따라서 본 논문에서는 GRU 알고리즘을 사용하여 토픽을 추출하였다.

Ⅲ. 실험결과

실험을 시작하기 전, 데이터 셋을 준비하였다. 데이터는 2019년 4월 23일부터 2019년 8월 23일까지의 트위터 뉴스 데이터를 사용하였다, 실험을 시작하고 가장 먼저 한 일은 데이터 전처리였다. 데이터 전처리는 카카오의 Khaii를 사용하여 형태소분석, 불용어(stop word)제거, 중복제거를 하였다. 그림 3은 지도학습을 한 후 테스트 케이스를 본 연구에서 만든 모델

에 입력 한 결과이다. 모델의 성능을 확인하기 위해 출력값은 SRC(본문), TGT(본문에서 추출된 토픽 - 타겟), PRD(모델에서 추출한 토픽)으로 이루어져 있다. 그림 3 을 보면 고유명사에 대한 형태소 분석이 완벽히 이루어지진 않았으나, 대체로 TGT과 유사한 PRD 값을 보임을 알 수 있다.

```
SRC :  ' 체포 영장 ' 김 장 겸 사장 화물 승강기 로 빠져나가 아
TGT :  김 / 장 / 겸
PRD :  김 / 장 / 겸

SRC :  ' 굳 은 표정 ' 이효성 방통 위원장
TGT :  이효성
PRD :  이효성

SRC :  잔뜩 긴장 하 ㄴ MBC 김 장 겸 사장
TGT :  MBC / 김 / 장 / 겸
PRD :  MBC / 김 / 장 / 겸

SRC :  퇴진 요구 받 는 MBC 김 장 겸 사장
TGT :  MBC / 김 / 장 / 겸
PRD :  MBC / 김 / 장 / 겸

SRC :  수행원 들 에게 둘러싸이 ㄴ MBC 김 장 겸 사장
TGT :  MBC / 김 / 장 / 겸
PRD :  MBC / 김 / 장 / 겸

SRC :  김장 겸 MBC 사장 체포 영장, 노조 ' 즉각 법정 세우 어 이 라 '
TGT :  김장 / 겸 / MBC / 체포 / 영장
PRD :  김장 / 겸 / MBC

SRC :  검찰 , 김장 겸 MBC 사장 체포 영장 발부 받 았 다
TGT :  김장 / 겸
PRD :  김장 / 겸 / MBC / 영장

SRC :  북한 " ICBM 장착 용 수소탄 시험 완전 성공 " <UNK>
TGT :  북한 / 수소탄
PRD :  ICBM / 용 / 수소탄

SRC :  휴대폰 <UNK> <UNK> 보다 인터넷 요금 인하 를 원하 ㄴ다
TGT :  휴대폰 / <UNK>
PRD :

SRC :  대안 학교 가 ㄴ다는 아들 교사 아빠 가 응원 하 ㄴ 이유
TGT :  대안 / 학교
PRD :  아들

SRC :  피 투성이 후배 폭행 하 고 사진 올리 ㄴ 중 3 " 건 방 지 어서 "
TGT :  폭행
PRD :  피 / 중

SRC :  18 회 <UNK> 와 ' 순수 ' 논쟁
TGT :  <UNK>
PRD :  18

SRC :  1969 년 오늘 - 베트 남 인민 의 ' 호 아저씨 ' 돌아가 다
TGT :  베트 / 남
PRD :  1969 / 베트 / 인민

SRC :  일 정부 " 북한 , 핵 실험 하 ㄴ 것 으로 단정 ... 강력 규탄 "
TGT :  북한 / 핵 / 실험
PRD :  북한
```

[그림 3] 학습된 모델 테스트 결과

Ⅳ. 결론

본 논문에서는 트위터 데이터 셋을 사용하여 GRU 알고리즘으로 키워드를 추출하는 모델을 구축하였다. 트위터 데이터셋에 Khaii 형태소 분석기를 사용하였다. 고유명사와 같이 출현 빈도 횟수가 낮은 것들은 불용어 처리를 하였다. 전처리된 트위터 데이터셋에 GRU를 사용하여 구축한 학습 모델에 학습 및 테스트 데이터로 활용하였다. 모델을 지도학습 후 테스트 데이터를 사용하여 [그림3]과 같이 지도학습 시 주어진 결과 값과 비슷한 결과를 도출하는 것을 확인하였다.

추후에는 뉴스 데이터셋을 수집하여 각 뉴스에 관한 키워드를 N-Gram으로 추출할 예정이다. N-Gram 키워드로 학습 데이터를 구축한 뒤 뉴스 데이터셋을 사용했을 경우에도 만족할 결과를 확인하는 연구를 진행할 것이다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음"(2015-0-00912)

참 고 문 헌

[1] 노준호. "지역별 토픽 추출을 위한 단어 군집화 방법." 국내석사학위논문 숭실대학교 대학원, 2014. 서울

[2]구범용. "사전 클러스터링을 이용한 Direct-LDA의 최적화." 국내석사학위논문 명지대학교, pp 6-8, 2006. 경기도

[3]김동욱, 이수원. (2017). 단어 유사도를 이용한 뉴스 토픽 추출. 정보과학회논문지, 44(11), 1138-1148.

[4] 이동준, 임유빈, 권태경. “형태소 기반 효율적인 한국어 단어 임베딩”. 정보과학회논문지 Vol.45, No. 5, pp. 444-450, May 2018.

[5] 우경진, 정수현. “문장 유형에 따른 한글 형태소 분석기 비교”. 한국소프트웨어종합학술대회 논문집, pp. 1388-1390, Dec, 2019.

[6] 김윤진. "딥러닝(Deep Learning)을 활용한 이미지 빅데이터(Big Data) 분석 연구." 국내박사학위논문 중앙대학교 대학원, 2017. 서울

[7] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares Holger Schwenk, Yoshua Bengio. “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation”, EMNLP 2014, Jun, 2014.

[8] Hochreiter S , Schmidhuber J . "Long Short-Term Memory" Neural computation : 1735-1780.

[9] Hochreiter S . "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions" International journal of uncertainty, fuzziness, and knowledge-based systems : 107-115.

[10]송경환. "GRU 기계학습 알고리즘을 이용한 델타 상관관계 프리페치 기법." 국내석사학위논문 인하대학교 대학원, 2018. 인천