합성곱 신경망 구조를 이용한 문서 범주 관련 키워드 추출

성수진 [1O], 방준성 [2$], 차정원 [1*]
창원대학교 [1], 한국전자통신연구원 [2]
{20153057 [O],jcha*}@changwon.ac.kr, hjbang21pp[$]@etri.re.kr

Keyword Extraction for Classification Using Convolutional Neural Network

Su-Jin Seong[1O], Jun-Seong Bang[2$], Jeong-Won Cha[1*]
Changwon National University[1], Electronics and Telecommunications Research Institute[2]

요 약

본 논문에서는 합성곱 신경망을 이용해 문서들의 범주와 연관성이 높은 키워드를 추출하는 방법을 제안한다. 합성곱 신경망에서 합성곱 층은 분류에 가장 영향을 많이 주는 층이다. 이 부분에서 가중치를 높게 받는 자질을 키워드로 추출한다. 논문 초록을 대상으로 한 실험에서 합성곱 신경망으로 추출한 키워드는 기존 방법으로 추출한 키워드에 비해 우수한 결과 보여주었다. 제안 방법은 분류 뿐만 아니라 문서 요약 등에서도 활용될 수 있을 것이다.

1. 서론

오늘날 디지털 데이터의 90%는 텍스트, 이미지, 동영상과 같은 비정형 데이터이며, 이러한 데이터로부터 가치를 도출하는 텍스트 분석 기술은 디지털 시대의 핵심역량으로 떠오르고 있다[1]. 방대한 양의 텍스트 문서들의 사용이 가능해지며, 유용한 정보를 찾아내는 것이 중요해졌다.

키워드는 다른 문서와 구분해 해당 문서를 대표할 수 있는 단어로, 정보 검색, 문서 분류, 요약 등에서 사용될 수 있다. 하지만 인간이 직접 키워드를 추출하는 것은 시간, 노력의 측면에서 비효율적인 일이다. 그렇기 때문에 자동으로 키워드를 추출할 수 있는 기술들이 연구되어왔다.

[2]에서는 TF-IDF 변형식을 이용하여 전자뉴스로부터 키워드를 추출하였다. [3]는 TF-IDF를 이용해 추출한 주제어를 기반으로 k-mean 클러스터링을 이용해 문서를 군집화하였다. fastText를 이용한 방법으로는 [4]가 문서 표현 방법으로 fastText를 사용하여 k-means 클러스터링 방법으로 문서를 군집화한 후 군집 키워드를 추출하는 방법을 제안하였다.

TF-IDF는 간단한 방법이지만 통계 기반으로 문서의 문장 수나 핵심 키워드의 출현 빈도가 낮으면 주요 키워드를 잘 찾아내지 못하는 문제가 있다.

fastText는 단어의 앞뒤 문맥을 반영하여 단어 임베딩을 연산한다. 이 경우 카테고리 단어와 유사도가 높은 단어를 구할 수 있지만 문장 내에서 카테고리 단어와의 거리가 먼 단어의 경우 키워드라도 유사도가 낮게 계산될 수 있다.

본 논문에서는 Convolutional neural networks (CNN)의 weights를 이용하여 문맥에 더하여 카테고리 정보를 반영함으로써 분류 문제에 도움을 주는 키워드로 구성된 사전을 생성하는 방법을 제안한다.

2. 제안 방법

CNN에서 convolutional layer는 입력 문장의 자질들에 대해 가중치를 계산한다. 그 후 max-pooling을 통해 각 필터에서 가장 높은 가중치를 가진 자질만 선택하여 카테고리를 분류한다. 모델은 예측 카테고리와 실제 카테고리 사이의 차이를 back-propagation을 통해 각 레이어에 전달하여 가중치를 업데이트한다. 이 과정에서 convolutional layer는 각 문장에서 카테고리를 분류하는데 있어 중요한 자질에 높은 가중치를 부여하도록 학습하게 된다.

우리는 convolutional layer의 가중치가 분류에 대한 자질의 중요도를 나타낸다는 점을 이용해, 가중치가 높은 단어를 키워드로 간주하고 추출한다.

이 때 오타와 같은 문서 내 오류에 유연하게 대처할 수 있도록 CNN을 음절 단위로 학습하여 해당 문서에 대한 각 음절의 가중치를 구한 후 단어를 이루는 음절들의 가중치를 조합하여 최종 중요도를 구한다. 본 논문에서는 단어를 이루는 모든 음절들의 가중치 합(CNN-SUM), 가장 높은 가중치(CNN-MAX), 가중치가 0이 아닌 음절들의 가중치 평균(CNN-

MEAN)을 중요도로 사용하였다.

제안 방법으로 추출된 키워드가 문서의 범주와 연관성이 높고 문서를 대표하기에 적합함을 증명하기 위해 TF-IDF와 fastText를 이용하여 키워드를 각 50개씩 추출하고 결과를 비교하였다.

3. 실험

3.1 학습 코퍼스

본 논문은 DBpia[5]에서 수집한 2,156개의 논문을 대상으로 논문의 초록을 이용해 논문의 분야를 분류하는 실험을 진행하였다. 이 때 논문의 연구 분야는 DBpia에서 설정한 주제 분류의 대분류를 따라 '공학', '농수해양학', '복합학', '사회과학', '예술체육학', '인문학', '의약학', '자연과학'으로 나눈다. 이 중 '공학'이 전체 960개로 가장 많고, '의약학'이 28개로 가장 적다. 전체 데이터 중 학습 데이터로 1,724개, 평가 데이터로 432개를 사용한다.

논문의 초록은 한글 초록을 사용하고, 키워드는 Espresso[6]를 이용한 형태소 분석 결과 중 의미형태소(체언(NN, NG, NR), 용언(VV, VA)-'하다', '되다' 제외-, 어근(XR), 외국어(SL))만을 사용한다. 논문 초록의 단어 빈도 수의 통계를 확인하였을 때 최대 빈도는 2050이었으며 평균은 3.005, 제 3 사분위수는 2였다. 이 통계 값을 기반으로 빈도 수가 2 이상인 단어만 키워드 후보로 사용한다.

3.2 실험 설정

3.2.1 TF-IDF

TF-IDF는 다음 식 1과 같이 계산된다.

d는 문서, D는 전체 문서, t는 단어, $f(w,d)$는 문서 d 내에서 나타나는 단어 w의 총 빈도이다. 이 때 다른 단어에 비해 과도하게 높은 출현 빈도를 보이는 단어의 영향을 줄이기 위해 sublinear TF scaling을 적용한다. $\{d \in D: w \in d\}$는 단어 w가 나타나는 문서 d의 개수이다.

$$tfidf(t,d,D) = \frac{1+log(f(w,d))}{1+\{d \in D: w \in d\}} \quad \text{식 1}$$

3.2.2 fastText

fastText는 텍스트 표현 및 분류를 위해 Facebook에서 개발한 라이브러리이다[7].

Epoch은 1000, window size는 5, 임베딩 차원은 50으로 설정하여 모델을 학습하였고, 코사인 유사도를 이용해 카테고리와 가장 유사한 단어를 추출한다.

3.2.3 CNN

키워드 추출에 사용할 레이어의 학습을 위해 논문 초록의 분야를 분류하는 CNN[8] 모델을 학습하였다. convolution filter는 3,4,5 크기의 커널을 사용하고 각 필터는 16개이다. 임베딩 차원은 50, batch는 32로 설정하였다. 이 모델이 초록을 분류하는 성능은 macro f1-score는 0.279, accuracy는 0.637이다. 이 모델을 기반으로 추출한 키워드의 카테고리는 입력 문서의 실제 카테고리를 기준으로 한다.

3.3 평가 방법

3.3.1 사전 매칭 분류

본 논문에서는 문서의 범주 정보를 반영하는 키워드를 추출하는 것을 목표로 한다. 따라서 추출된 키워드와 카테고리 사이 연관성을 확인하기 위해 카테고리별로 추출된 키워드 50개의 출현 빈도를 이용하여 평가 문서를 분류한다.

예측 카테고리는 식 2와 같이 계산된다. $tf(w_{ki},d)$는 입력 문서 d에서 카테고리 k의 i번째 단어 w_{ki}의 출현 빈도이고, $cf(w_{ki},C)$는 전체 카테고리 키워드 리스트 C 중 w_{ki}가 출현하는 카테고리 수이다.

예측 카테고리 $\hat{y}$와 실제 초록의 카테고리를 비교하여 키워드 리스트의 분류 성능이 높을 수록 범주의 특징을 적절히 반영하였다고 판단한다.

$$\hat{y} = argmax_k \left(\sum_i \frac{tf(w_{ki},d)}{cf(w_{ki},C)} \right) \quad \text{식 2}$$

3.3.2 Fleiss' Kappa

Fleiss' Kappa는 두 명 이상의 평가자 사이의 합의 또는 한 명의 평가자에 대한 신뢰도를 평가할 때 사용하는 통계적 방법이다.

추출된 키워드들은 각 카테고리에 적합한 키워드인지 다수의 평가자들에 의해 1에서 5 사이의 값으로 점수가 매겨진다. Fleiss' Kappa의 결과값에 따라 ≤ 0 부족한, 0.0-0.2 약간, 0.21-0.4 조금 큰, 0.4-0.6 중간의, 0.61-0.8 상당한, 0.81-1.0 거의 완벽한 일치를 의미한다[9].

3.4 실험 결과 및 분석

각 모델별 사전 매칭 분류 결과와 Fleiss' kappa 결과는 표 1과 같다. Human Score는 8개의 카테고리에 대해 추출된 50개의 키워드에 대한 각 평가자들의 평균 점수를 구한 결과이다. k는 Fleiss' kappa 값을 나타낸다.

CNN-SUM과 CNN-MEAN을 제외하고 모두 약간의 일치를 보인다. 사전 매칭 분류 결과 단어를 이루는 음절의 가중치의 최고 값을 중요도로 두었을 때의 f1-score가 0.2986로 가장 높았고 accuracy는 0.375로 두 번째로 높으며, 또한 평가자들의 일치도 가장 높았다. CNN-MEAN 모델은 accuracy가 0.3935로 가장 높고, f1-score가 0.2791로 두 번째로 높았다.

표 1 사전 매칭 분류 결과 및 Fleiss' kappa 결과 성능표

Model	Macro F1-Score	Accuracy	Human Score	k
TF-IDF	0.2086	0.2014	1.439	0.226
fastText	0.1009	0.1528	1.462	0.247
CNN-SUM	0.2646	0.3634	**1.993**	0.172
CNN-MAX	**0.2987**	0.3750	1.775	0.228
CNN-MEAN	0.2791	**0.3935**	1.424	0.203

표 2 '인문학' 키워드 리스트 상위 9개

TF-IDF	fastText	CNN-SUM	CNN-MAX	CNN-MEAN
니체	가르치	크레쯔만	개인	철학
예술	인격형성	절충주의	정치공동체	개념
예술철학	직	훌륭	왈쩌	삶
철학	가치실현	개체사이의분볓	떠들썩	인간
창조	전화	엄숙주의	초대교회	행복
의지	변모	떠들썩	도덕철학	어떻
논의	경험	자유주의	칸트	탐구
삶	어선	뒷받침	윤리	패러다임
탐구	더불	데모크리토스	헤겔	의지

표 2는 모델별 키워드 리스트 중 '인문학' 카테고리의 키워드 예이다. 추출된 키워드의 예를 보면 TF-IDF의 경우, 해당 카테고리와 의미 상으로 연관이 있지만 문서 전체에서 일반적으로 사용될 수 있는 단어가 추출되는 경향을 보인다. 이와 비교해 CNN 모델로 추출한 키워드들은 카테고리와 연관성이 높은 단어이지만, 해당 카테고리 내 문서 사이에서도 구별이 되는 단어들임을 알 수 있다. CNN-SUM의 경우 두드러지게 구별되는 단어를 추출하였고, 평가자들의 평균 점수가 1.993으로 가장 높지만 평가자들의 일치도 k가 가장 작다. 이는 이 단어들이 평가자들의 일치를 보지 못하고 있음을 나타낸다. 사전 매칭 분류에서는 세 번째로 높은 성능을 보였고, 재출현 확률이 낮은 키워드가 추출되었기 때문으로 판단된다. 예로 '자유주의'는 '인문학'에서 18개의 문서, '엄숙주의'는 1개의 문서에서 출현한다.

CNN-MAX의 경우 해당 카테고리에 주로 나타나지만 CNN-SUM에 비해 재출현 확률이 높은 단어들이 추출되었다. 예를 들어 '칸트'는 39개의 '인문학' 문서에서 나타나고, '헤겔'은 23개의 문서에서 나타난다. 이에 평가자들의 점수는 1.775로 두 번째로 높은 점수를 얻었으며 분류 성능 또한 우수하였다.

4. 결론

본 논문에서는 문맥 정보를 반영하며 문서 범주와 연관성이 높은 단어 사전을 구축하는 것을 목표로, CNN 분류 모델의 학습된 가중치를 이용해 키워드를 추출하는 방법을 제안하였다. CNN의 convolutional layer는 입력 자질에 가중치를 부여하고 max-pooling은 그 중 분류에 가장 중요한 자질을 선택한다. 즉 분류 성능을 높이는 중요한 자질의 가중치를 높이도록 학습된다. 이렇게 학습된 가중치를 이용하여 입력 문서로부터 키워드를 추출하였을 때 TF-IDF나 fastText에 비해 다른 범주와 명확히 구별되는 단어들이 추출되었다.

더하여, 단일 문서에서 문서의 내용을 대표하고 다른 문서들과 구분하기 위해서는 문서의 문맥 상 유의미하며 유니크한 단어를 키워드로 추출하는 것이 효과적일 수 있지만, 문서들을 정해진 카테고리로 분류하기 위해서는 추출된 단어의 재출현 가능성 또한 고려해야하였다.

제안 방법 중 CNN에 의해 결정된 음절들의 가중치의 최댓값을 해당 단어의 중요도로 사용하는 것이 이러한 조건을 충족시켜 높은 분류 성능을 보였고 사람의 평가에서도 약간의 일치 범위에서 높은 성능을 보임을 확인하였다.

또한 CNN을 사용하여 키워드를 추출하였을 때 일반적으로 나타나지않는 키워드로 추출되었다. 이를 요약문제에 활용할 경우 요약문의 다양성을 높일 수 있을 것으로 기대된다.

Acknowledgement

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[No.2018-0-00440, 위험 상황 초기 인지를 위한 ICT 기반의 범죄 위험도 예측 및 대응 기술 개발]

참 고 문 헌

[1] 박현진, "SAS, 비정형데이터에서 텍스트 분석 기술은 디지털 시대의 핵심역량", 인공지능신문, http://www.aitimes.kr/news/articleView.html?idxno=12114, accessed April 28, 2019

[2] 이성직, 김한준, "TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법", 한국전자거래학회지, 14권, 4호, pp. 59-73, 2009

[3] 이윤수, 길준민, 이종혁, They Pheaktra, "키워드 군집화를 이용한 연구 논문 분류에 관한 연구", 한국정보처리학회논문지 소프트웨어 및 데이터 공학, 7권, 12호, pp. 477-484, 2018

[4] 박원상, 양예진, 이수진, 이한빛, 연종흠, 이상구, "fastText를 이용한 문서 군집화 및 군집의 특성을 고려한 군집 키워드 추출", 한국정보과학회 학술발표논문집, pp. 820-822, 2018

[5] 'DBpia', https://www.dbpia.co.kr/, accessed April 21, 2019

[6] 'Espresso', http://air.changwon.ac.kr/~airdemo/Espresso/, accessed April 28, 2019

[7] Armand Joulin & Edouard Grave & Piotr Bojanowski & Tomas Mikolov, 'Bag of Tricks for Efficient Text Classification', arXiv:1607.01759 [cs.CL], 2016

[8] Kim Yoon, 'Convolutional neural networks for sentence classification', arXiv preprint at arXiv: 1408.5882, 2014.

[9] 박창언, 김현정, "체계적 문헌고찰에서 평가자 간의 신뢰도 측정", Hanyang Med Rev, 35권, 1호, pp. 44-49, 2015