문단 간 유사도와 로지스틱 회귀 분류를 통한 기사 내 부적합 문단 검출 시스템

Detecting Improper Paragraphs in a News Article Using Logistic Regression Classification and Inter-class Similarity

저자 (Authors)
김규완, 신현주, 김선진, 문경득, 이현아
KyuWan Kim, HyunJu Shin, SunJin Kim, KyoungDuek Moon, HyunAh Lee

출처 (Source)
한국정보과학회 학술발표논문집 , 2017.12, 1873-1875(3 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07322713

APA Style
김규완, 신현주, 김선진, 문경득, 이현아 (2017). 문단 간 유사도와 로지스틱 회귀 분류를 통한 기사 내 부적합 문단 검출 시스템. 한국정보과학회 학술발표논문집, 1873-1875

문단 간 유사도와 로지스틱 회귀 분류를 통한 기사 내 부적합 문단 검출 시스템

김규완○, 신현주, 김선진, 문경득, 이현아
금오공과대학교 컴퓨터소프트웨어공학과
rla9826@naver.com, dotcomehe@naver.com, junnis0123@naver.com,
ans5097@hanmail.net ,halee@kumoh.ac.kr

Detecting Improper Paragraphs in a News Article Using Logistic Regression Classification and Inter-class Similarity

KyuWan Kim○, HyunJu Shin, SunJin Kim, KyoungDuek Moon, HyunAh Lee
Dept. of Computer Software Engineering, Kumoh National Institute of Technology

요 약

최근 들어 뉴스의 본질을 해하는 속칭 가짜 뉴스가 만연하고 있어 이를 빠르게 검출하기 위한 요구도 늘어나고 있다. 텍스트마이닝을 이용한 기존의 부적합 문단 검출 시스템에서는 1)유사도 연산에 상당한 시간이 소요되고, 2)연산 결과의 정확도가 낮으며 3)시스템의 전체적인 정확도가 저하되는 문제가 있었다. 본 논문에서는 위 문제를 해결하기 위하여 Word2vec 대신 Doc2vec을 활용하여 문단 간 유사도를 측정하고, 학습데이터를 늘린 분류 분포도 모델을 적용하는 부적합 문단 검출 모델을 제안한다. 제안 모델은 시스템 속도가 18배 상향되었고, 분포도 모델의 정확도가 98.1%로 개선되었으며, 전체 시스템 정확성 또한 92.07%로 향상되었다.

1. 서 론

최근 들어 SNS와 스마트기기의 발전으로 온라인을 통한 뉴스 배포가 용이해지면서 악의적인 문단 조작이나 광고성 문단을 포함하는 속칭 가짜 뉴스가 대량 생산되고 있다. 이러한 문제 해결을 위해 제안된 기존의 텍스트마이닝을 이용한 부적합 문단 검출 시스템[1]에서는 문단 간 유사도와 문단의 뉴스 분류 분포도에 기반 하여 부적합 문단을 검출하였다. 문단 간 유사도는 Word2vec에 기반 하여 얻어지며, 분류 분포도는 로지스틱 회귀를 적용하여 각 문단의 뉴스 분류를 결정한 뒤 이질적인 분류로 파악된 문단에 낮은 점수를 부여한다. 그러나 이 모델은 시스템 속도와 유사도 모델 정확도 추출에서 낮은 성능을 보였다. 본 논문에서는 기존 모델의 속도 개선 및 정확도 개선을 위하여 유사도 모델 메커니즘에 Word2vec대신 Doc2vec을 적용한 새로운 부적합 문단 검출 시스템을 제안한다.

2. 관련 연구

CNN모델 중 하나인 Word2vec은 입력한 말뭉치의 문단에 있는 단어와 인접 단어의 관계를 이용하여 단어의 의미를 학습한다. Word2vec의 학습 방법은 CBOW, Skip-gram의 두 종류가 있다. CBOW(Continuous Bag Of Words)방식은 주변 단어가 만드는 맥락을 이용해 타겟 단어를 예측하고, Skip-gram은 한 단어를 기준으로 주변에 올 수 있는 단어를 예측한다[2]. 이에 반하여, Doc2vec은 Word2vec을 문단, 단락 또는 전체 문서와 같이 더 큰 텍스트 블록에 대한 연속 표현을 학습하도록 수정한 모델로 단어와 단어가 아닌 문단과 문단, 문서와 문서 등 더 큰 벡터에서 속도와 정확도가 높은 장점이 있다[3].

부적합 문단을 추출하기 위한 기존의 텍스트마이닝을 이용한 연구에서 분포도 모델은 전처리를 거친 뉴스 데이터를 TF-IDF 가중치 알고리즘을 적용하여 임베딩한다. 임베딩된 벡터와 레이블을 Scikit-learn library에서 제공하는 로지스틱 회귀 분류(Logistic regression)[4] 을 이용하여 학습시킨다. 학습된 모델은 n개 문단을 입력 값으로 받고, 각 문단의 레이블을 판단한다. 문단의 분포도는 해당하는 레이블의 총 개수를 전체 문단의 수로 나눈 값을 받는다. 예를 들어 4개의 문단 중 3개의 문단이 같은 분류로 판단되면, 세 개의 3/4을, 나머지 한 문단은 1/4을 분포도로 얻어, 이질적인 분류의 문단이 낮은 점수를 받게 한다.

유사도 모델은 komoran형태소 분석기로 명사만 추출된 뉴스 데이터를 Word2vec으로 학습시킨다. 학습된 모델은 각 문단에 존재하는 명사 벡터 사이의 유사성을 기준으로 문단의 유사도를 계산한다. 유사도는 입력받은 각 n개의 문단 모든 명사에 대해 대상 문단을 제외한 다른 문단 명사와의 유사도를 구하고, 얻어진 유사도 값들을 n으로 나눠 각 문단의 평균 유사도를 구한다. 즉, 유

사도 연산 소요 시간은 문장 길이에 비례한다. 시스템에서는 최종적으로 유사도와 분포도를 곱한 값으로 결정하여, 가장 높은 값과 30%이상의 차이가 나면 부적합 문단으로 판별한다.

기존 연구 모델에서 Python Flask를 통해 웹 서비스를 제공하였으나 유사도 계산에 많은 시간이 소요되어 서비스 품질이 저하되었다. 본 논문에서는 기존 시스템에서의 복잡한 연산과정을 가진 유사도 모델을 개선하기 위해 Doc2vec을 이용하고자 한다. 또한, 기존 Word2vec을 이용하여 문단의 단어 하나하나를 비교하는 방법보다 Doc2vec 학습 모델을 통해 문단 전체를 비교함으로써 유사도 모델의 정확도를 개선하고자 한다.

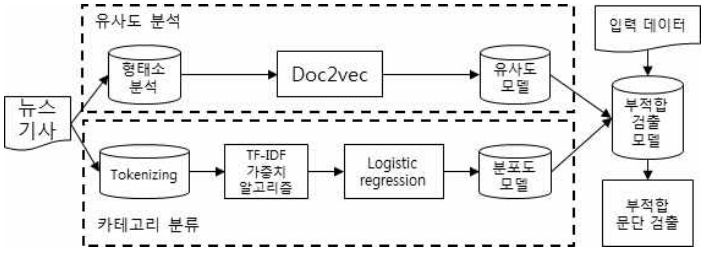
3. 제안 시스템

본 논문에서는 Doc2vec을 통해 학습시킨 유사도 모델을 이용하는 부적합 문단 검출 시스템을 제안한다. 분포도 모델은 기존과 동일하게 공백단위로 토큰화 전처리를 거친 뉴스 데이터를 TF-IDF 가중치 알고리즘으로 임베딩한 벡터와 레이블을 Logistic regression에 적용하여 학습시킨다.

기존 모델에서는 분포도와 유사도를 모든 입력 값에 적용하여 문단 간 연관도를 계산하는 반면, 제안 모델에서는 모든 문단의 카테고리 분포도가 0.5 이하이거나 1일 경우에만 유사도 모델을 적용한다. 즉, 동일한 카테고리 범주 내에서의 부적합 문단의 판단에는 유사도 모델이 이용되는 것이다. 이는 카테고리가 상이한 부적합 문단일 경우 유사도 모델이 연관도에 큰 영향을 미치지 못하는 것에서 착안하였다. 제안 시스템의 구조는 그림1과 같다.

학습 데이터 셋은 유사도 모델의 경우 기존 모델과의 성능 비교를 위해 동일하게 2017년 8월~ 2007년 3월 일자까지 조선일보에서 수집한 뉴스 데이터로 구성한다. 뉴스 데이터는 각 카테고리 별 5만 건의 뉴스 기사로 총 30만 건이다. 분포도 모델의 경우 학습 데이터 셋은 동일 기간의 조선일보와 동아일보에서 수집한 총 60만 건의 뉴스 데이터로 구성한다.

분포도 모델에서는 기존 시스템과 같은 (0)스포츠, (1)정치, (2)경제, (3)사회, (4)연예/방송, (5)오피니언/칼럼/사설로 나누어 0~5의 식별자를 부여한 레이블을 함께 학습한다. 학습된 분포도 모델은 문단을 입력하면 해당 문단의 카테고리를 반환한다. 시스템은 반환받은 레이블로 총 개수를 전체 문단의 수로 나눈 분포도를 계산한다.



[그림 1] 시스템 구조도

또한, 분포도 모델의 경우 정확도가 학습 데이터양에 비례했다. 기존 모델에서는 30만 건의 학습 데이터로 87%의 성능을 보여주었으나 60만 건의 학습 데이터는 98.1%까지 정확도가 상승했다. 제안 시스템에서는 60만 건의 학습 데이터로 학습한 분포도 모델을 사용한다.

기존 연구에서는 Word2vec을 이용하여 학습시킨 유사도 모델을 이용하였는데, 해당 모델은 문단에 포함된 단어 하나하나의 유사도를 비교하여 총 문단의 유사도를 구하는 연산 과정을 거쳤다. 이로 인해 연산 시간이 평균 3분 가까이 소요되었으며, 정확도도 기대에 미치지 못하였다.

본 논문에서는 기존 모델의 문제점을 개선시키기 위해 유사도 모델을 Doc2vec으로 학습시켰다. Word2vec과 Doc2vec은 형태소 분석기를 거쳐 뉴스 전체의 명사를 추출하여 명사단위로 학습을 시키는 형태는 동일하나 유사도를 계산하는 과정이 다르다. Word2vec의 유사도 연산은 입력 값이 단어로 이루어지며, 결과 값도 단어와 단어 사이의 유사도 점수로 계산된다. 그러나 Doc2vec은 입력 값이 문단으로 이루어지고, 문단과 문단간의 유사도 연산을 수행할 수 있다. 또한, Word2vec을 통한 문단 유사도 계산의 경우 문단을 단어로 분리하여 단어 유사도의 평균을 구하는 비효율적인 과정을 거치는 반면, Doc2vec을 통한 문단 유사도 계산에서는 그 과정이 생략된다. 그러므로 속도와 정확도가 향상되는 결과를 얻을 수 있다. 그리고 제안 시스템에서는 기존 연구에서는 사용하지 않던 뉴스 제목과 문단과의 유사도를 구하여 각 문단에 더하여 정확도를 개선하고자 한다.

최종적으로 분포도와 유사도의 곱으로 문단 간 연관도를 구한다. 문단 간 연관도가 제일 높은 문단과 10% 이상의 차이를 보이면 부적합 문단이라고 판단한다.

4. 성능 실험 및 결과

4.1 로지스틱 회귀 분류학습 데이터양에 따른 정확도

본 논문에서 분포도 모델인 TF-IDF를 이용한 로지스틱 회귀 분류에 학습데이터의 양을 각각 5000, 10000, 50000, 100000개에 대해 테스트를 진행했으며 정확도가 78.6%, 82.4%, 87.0%, 98.1%로 점점 증가했다.

4.2 기존 연구와 성능 비교

시스템 정확도를 평가하기 위하여 조선일보에서 2017년 9월 뉴스를 수집한 뒤 부적합 문단이 포함되지 않은 1만 건의 데이터와 임의로 타 카테고리 기사의 한 문단을 포함시킨 데이터와 같은 카테고리 내에서도 부적합한 문단을 포함시킨 1만 건의 데이터로 테스트 데이터 셋을 구축했다. 부적합 문단이 포함된 뉴스는 43214개, 포함되지 않은 뉴스는 43218개의 문단으로 구성된다.

제안된 알고리즘의 성능 평가에는 검출 분야에서 사용되고 있는 평가 방법 중 하나인 정확성(Accuracy)을 이용한다. 정확성은 아래 수식을 통하여 계산한다.

$$\text{정확성(Accuracy)} = \frac{TP+TN}{TP+TN+FP+FN}$$

평가에서는 아래 각 모델에 대한 성능 비교를 수행하였다.

(가) Word2vec + Logistic regression (기존 모델)
(나) Doc2vec + Logistic regression
(다) Doc2vec + Logistic regression + 제목 유사도

아래는 각각 (가), (나), (다) 모델에 대하여 진행한 성능 비교 결과를 보인다. 결과에서 볼 수 있듯이, 제안한 모델 (다)모델이 가장 높은 성능을 보이는 동시에 가장 빠른 결과를 얻었다.

		label		정확도	평균 속도
		True	False		
(가)	True	TP : 38896	FP : 4422	89.88%	2분24초 이내
	False	FN : 4322	TN : 38792		
(나)	True	TP : 39458	FP : 4192	90.79%	15초 이내
	False	FN : 3760	TN : 39022		
(다)	True	TP : 39760	FP : 3389	92.07%	15초 이내
	False	FN : 3458	TN : 39825		

5. 결 론

본 논문에서는 부적합 문단 검출 시스템의 성능 향상을 위해 학습 데이터양을 증가시키고 Doc2vec을 활용하였다. 학습 데이터양을 증가시킴으로써 카테고리 분류를 위한 로지스틱 회귀 분류 모델의 정확도는 87.0%에서 98.1%로 향상되었다. 문단 간 유사도를 구하기 위해 Word2vec 대신 Doc2vec을 사용함으로써 시스템 속도와 정확도를 개선할 수 있었다. 이에 따라 전반적으로 시스템의 성능이 개선되었으며, 제안 모델을 통해 최근 들어 이슈가 되고 있는 악의적으로 뉴스 본문을 조작하여 뉴스 본질을 흐리는 기사를 검출 해낼 수 있을 것이다. 또한, 뉴스뿐만 아니라 자기소개서, 논설문 등 부적합 문단이 포함될 수 있는 시스템에 적용 가능 할 것이다. 향후 연구로는 제목과 내용의 불일치 검출 연구를 통해 낚시성 제목이나 의도적으로 왜곡된 제목을 검출하고자 한다. 본 논문에서 제안된 시스템과 함께 이용한다면 가짜 뉴스 검출 시 더 정밀한 시스템이 될 수 있을 것이다.

참고문헌

[1] 김규완, 신현주, 김선진, 이현아, “텍스트마이닝을 이용한 기사 내 부적합 문단 검출 시스템“, 제29회 한글 및 한국어 정보처리 학술대회 논문집(2017년), 2017

[2] 이진욱, 유국현, 문병민, 배석주, “감성분석과 Word2vec을 이용한 비정형 품질 데이터 분석”, 한국품질경영학회, <품질경영학회지> 45권1호 (2017), 2017

[3] Quoc Le, Tomas Mikolov, “Distributed Representations of Sentences and Documents, 1600 Amphitheatre Parkway, Mountain View, CA 94043

[4] “로지스틱 회귀 분류”, IBM Knowledge Center