임베딩 기반의 비정형 문서 핵심 영역 식별

박민지[1], 황영준[1], 박병훈[1], 신수연[2], 이치훈[1*]
[1]티쓰리큐(주), [2]한양대학교(서울)
smh0924@t3q.com, hyjun0103@t3q.com, warmpark@t3q.com,
shinsy@hanyang.ac.kr, chihoon.lee@t3q.com

A method based on embedding to detect core regions in unstructured document

Min Ji Park[1], Yeong Jun Hwang[1], Byung Hoon Park[1],
Sooyeon Shin[2], Chi hoon Lee[1]
[1]T3Q(주), [2]Center for Creative Convergence Education, Hanyang University(Seoul)

요 약

기업의 운영에 있어서 기업의 핵심 정보가 유출되지 않도록 관리하는 것은 매우 중요하다. 따라서, 사내에서 유통되는 문서들에 대해 핵심적인 정보가 사외로 유출되지 않도록 관리하고 추적하는 것은 필수적이다. 특히, 데이터가 구조화되지 않고, 다양한 형식으로 구성되어있는 비정형 문서 내에서 핵심 정보를 식별하는 것은 기술적으로 어려움이 존재한다. 본 논문에서는 YOLOv8을 사용하여 비정형 문서 내에서 영역을 식별하고, 자연어 처리 모델인 Word2Vec을 사용하여 비정형 문서 내에서 핵심 내용을 식별한 후 이를 시각화함으로써 사내에서 유통되는 비정형 문서 내의 핵심 정보를 식별하고 추적하는 방법을 제안하였다.

1. 서론

산업기밀 유출은 기업의 경쟁력과 비즈니스에 심각한 피해를 초래한다. 따라서, 이러한 기밀 유출을 막기 위해 사내에서 유통되는 문서들에 대해 핵심 정보가 사외로 유출되지 않도록 방지하는 것은 매우 중요하다. 그러나 다양한 양식의 문서에서 핵심 정보를 식별하는 것은 여전히 기술적으로 어려움이 존재한다[1].

문서의 내용을 분석할 때 크게 정형 문서와 비정형 문서로 구분한다. 정형 문서란, 구조와 형식이 미리 정해져 있는 양식에 내용이 채워져 있는 문서를 의미하며, 비정형 문서는 구조와 형식이 일정하게 정해져 있지 않은 형태의 문서를 의미한다. 이러한 차이로 인해서 정형 문서는 상대적으로 간단한 기술을 통해 정보를 추출할 수 있지만, 비정형 문서에서 정보를 추출하는 것은 간단하지 않다[2][3].

이러한 배경으로 본 논문은 사내에서 사용하고 관리하는 문서 중, 특히 비정형 문서에서 핵심 정보를 담고 있는 영역을 식별하는 모델을 제안하고자 한다. 문서 데이터에서 핵심 영역은 사외로 유출되지 말아야 하는 내용을 포함한 영역을 의미하며, 본 연구에서는 테이블로 제한하고 연구를 진행한다.

논문의 구성은 다음과 같다. 2장에서는 연구에 사용한 관련 기술에 대해 설명한다. 3장에서는 전체적인 제안 방법과 각 단계별 처리 내용을 설명하고, 4장에서는 결론과 향후 연구에 관해서 서술한다.

2. 관련 기술

2.1 OCR

광학 문자 인식(Optical character recognition; OCR)은 이미지 상의 텍스트를 기계가 읽을 수 있는 텍스트 포맷으로 변환하는 기술이다. OCR은 이미지 상의 텍스트를 잘 인식할 수 있도록 이미지를 보정하는 전처리 단계, 이미지에 존재하는 텍스트 영역을 식별하는 문자 검출 단계, 텍스트 영역 안에 문자를 인식하는 문자 영역 단계와 같이 3단계의 처리 과정으로 진행된다[4].

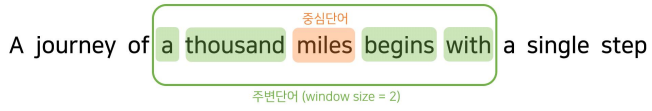
2.2 Word2vec

자연어 처리 기법에는 단어의 의미를 벡터값으로 수치화하는 워드 임베딩 방식이 있다. 단어를 벡터로 바꾸는 방법에는 희소 표현(Sparse

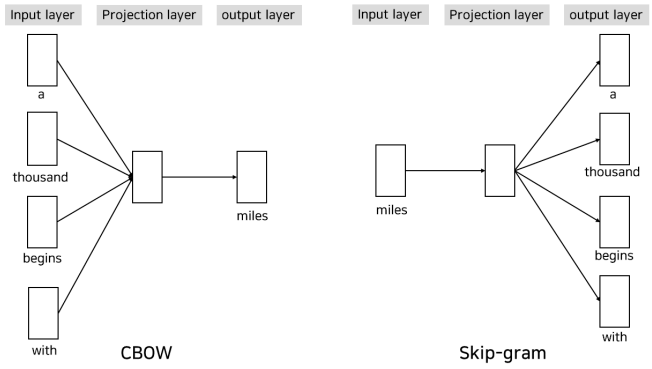
* 교신저자

Representation)과 분산 표현(Distributed Representation) 두 가지 방법이 있다. 희소 표현은 N개의 단어가 있다고 가정하면, N개의 단어를 표현하기 위해 N차원의 벡터를 만들고 단어의 속성을 표현하는 요소가 1의 값을 가지고, 나머지는 0의 값을 가지는 표현 방법이다. 간단하게 말하자면, one-hot encoding으로 만들어진 벡터 표현이다. 전통적으로 사용된 방법이긴 하지만, 각 단어 벡터 간 유사성을 표현할 수 없다는 단점이 존재한다. 이러한 단점을 보완한 방법이 분산 표현 방식이다. 분산 표현은 희소 표현 방식처럼 각각의 속성을 독립적인 차원에 나타내지 않고, 사용자가 정한 차원으로 대상을 대응시켜 표현한다. 임베딩한 벡터는 모든 차원이 값을 가지고 있는 벡터로 표현된다. 분산 표현 방식은 적은 차원으로 단어를 표현할 수 있고, 단어와 단어의 관계를 표현할 수 있게된다[5].

워드 임베딩 모델인 Word2Vec은 분산 표현 기법을 사용하기 때문에 각 단어의 의미를 고려한 특성 벡터를 추출할 수 있다[6]. Word2vec은 크게 CBOW(Continuous Bag Of Words) 방식과 Skip-gram으로 학습한다. 두 방식 모두 앞뒤로 주변 단어들을 몇 개씩 볼건지를 결정하는 window size를 사용하여 학습한다. CBOW 방식은 문맥을 바탕으로 단어를 예측하는 방식이다. 반대로, Skip-gram 방식은 단어를 바탕으로 문맥을 예측하는 방법이다. Skip-gram 방식은 역전파 관점에서 CBOW 방식보다 더 많은 학습이 일어나기 때문에, 성능이 더 우수한 것으로 알려져 있고, 많은 연구에서 Skip-gram 모델을 기본 채택하여 활용한다. 본 연구에서도 Skip-gram 방식을 사용하였다[7].



(그림 1) Word2Vec의 window size



(그림 2) Word2Vec의 CBOW와 Skip-gram

2.3 YOLOv8

YOLO는 실시간 객체 검출을 목표로 하는 딥러닝 기반의 알고리즘으로, 이미지나 비디오에서 여러 객체의 위치와 클래스를 신속하게 식별할 수 있는 모델이다. YOLOv8은 YOLO 모델 중에서 가장 최근에 출시된 SOTA(State-Of-The-Art) 모델로서 가장 우수한 성능을 보인다. 본 연구에서는 문서 내에 포함하고 있는 단위 테이블을 식별하기 위해 YOLOv8 detection model을 사용하였다.

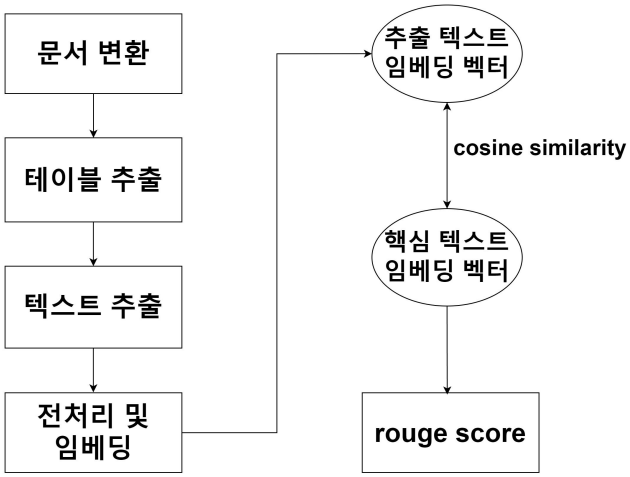
2.4. Cosine Similarity

코사인 유사도는 두 벡터 간의 코사인 각도를 이용하여 구할 수 있는 두 벡터의 유사도를 측정하는 계산적 방법이다. 0에서 1 사이의 값으로 결과가 산출되며, 1에 가까울수록 두 벡터가 유사함을 의미한다. 본 논문에서는 OCR을 통해 추출한 텍스트의 임베딩 벡터와 사전 정의된 핵심 텍스트의 임베딩 벡터 간에 얼마나 유사한지를 측정하기 위해서 Cosine Similarity를 사용하였다.

3. 제안 방법

3.1 영역 검출 방법 설계

비정형 문서 내 핵심 영역 식별에 대한 전체 프로세스는 (그림 3)와 같다.



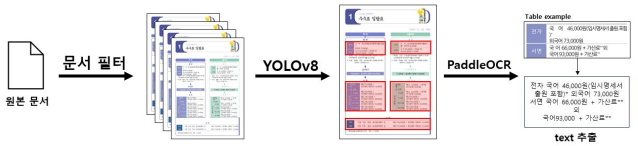
(그림 3) 비정형 문서 내 핵심 영역 식별 프로세스

3.2 후보 영역 검출

비정형 문서 내 핵심 영역 식별하기 전 단계로 후보 영역 검출을 진행한다. 입력값으로 원본 문서 파일을 받아, 문서 필터를 사용하여 문서의 페이지

별 PNG 파일을 추출한다. 추출된 PNG를 대상으로 YOLOv8을 사용하여, 단위 테이블을 식별하고, crop 이미지로 저장한다.

테이블 식별 모델은 문서에 존재하는 테이블 3,832개를 학습 데이터를 사용하여 학습시켰고, mAP50 수치는 0.942를 달성하였다. 마지막으로 저장된 crop 이미지를 대상으로 OCR을 진행하여 crop 이미지에 포함된 텍스트를 추출한다.



(그림 4) 후보 영역 검출 과정 시각화

3.3 텍스트 추출

대표적인 OCR 모델은 Tessaract, EasyOCR, PororoOCR, PaddleOCR이 있으며, 1차적으로 추출된 테이블 후보 영역에서 텍스트를 추출하는 실험을 하였다. 그리고 그 성능을 비교하였을 때, 속도와 성능 측면에서 PaddleOCR이 가장 우수함을 확인할 수 있었고, 따라서 본 연구에서는 PaddleOCR을 사용하였다.

<표 1> OCR 성능 비교 결과 표

	Tesseract	EasyOCR	PororoOCR	PaddleOCR
속도 (장 당)	3	2	4	1
성능	4	2	3	1
사용성	2	1	3	1

3.4 임베딩 벡터

추출된 텍스트를 임베딩 벡터로 변환하는 과정이다. 임베딩 모델의 입력값으로 OCR에서 추출된 텍스트를 받는다. 추출된 텍스트는 숫자 및 불용어를 제거하고, 명사만 추출한다. 연구에 사용된 데이터는 개조식 문장이 대부분이기 때문에 명사만 추출하여도 그 의미를 충분히 반영할 수 있다고 판단하였다. 추출한 명사들은 워드 임베딩 모델인 Word2Vec을 통해 벡터화한다.

연구에 사용한 데이터의 내용을 반영하는 symantic vector를 추출하기 위해 문서 데이터의 텍스트 전체를 말뭉치로 만들고, Word2Vec 모델을 학습하였다. Word2vec을 제외한 FastText, Glove 등 다른 단어 임베딩 벡터들도 성능을 비교하였으나, 핵심 내용을 식별하는 본 연구에서는 Word2vec이 가장 성능이 우수함을 확인할 수 있었다.

3.5 핵심 영역 식별 평가

미리 정의되어 있는 핵심 카테고리의 임베딩 벡터와 후보 영역에서 추출한 임베딩 벡터를 코사인 유사도 비교를 진행하고, threshold 이상인 것들만 선별한다. 그 후에, 코사인 유사도의 threshold를 만족한 값들을 대상으로 Rouge Score를 적용한다.

본 논문에서는 핵심 텍스트 식별의 성능을 향상시키기 위해서, 텍스트 요약 모델의 성능 평가 지표로 사용하는 Rouge Score[8]를 도입하고 하나의 단어 기준으로 측정하는 unigram 방식의 ROUGE-1을 사용하였다. 핵심 카테고리 별로 사전에 정의한 10개의 핵심 키워드가 추출한 텍스트에 얼마나 포함되어 있는지를 평가하고 비정형 문서 내 핵심 영역을 식별한다. 코사인 유사도와 마찬가지로 threshold 이상인 값들만 선별하여 최종적으로 핵심 영역을 식별하게 된다.

모든 단계가 끝나면, 비정형 문서 내 핵심 영역이 존재하는 페이지를 (그림 5)와 같이 시각화하여 출력한다. 상단에는 문서에 관련된 정보인 문서명과 문서 유형, 페이지 번호, 해당하는 핵심 카테고리, 코사인 유사도를 출력한다. 중간에는 핵심 영역이 포함된 페이지의 썸네일을 출력한다. 핵심 영역은 바운딩 박스로 하이라이트 하여 표시하고, 상단과 마찬가지로 핵심 영역 관련 정보(핵심 카테고리, 코사인 유사도)를 바운딩 박스 위에 시각화한다. 마지막으로 하단에는 핵심 영역의 crop image를 확대하여 출력해 실제로 핵심 영역에 어떤 내용이 있는지 볼 수 있도록 하였다. 연구에 사용된 데이터는 외부에 공개가 불가능하기 때문에 공공데이터포털에서 샘플 데이터를 가져와 시각화된 모습을 재현하였다.

(그림 5) 비정형 문서 내 핵심 영역 식별 시각화

4. 결론 및 향후 연구

본 논문은 비정형 문서 내에서 핵심 영역을 식별하기 위한 방법을 제안했다. YOLOv8, PaddleOCR, Word2Vec, Cosine Similarity, Rouge Score를 활용하여 비정형 문서에서 테이블을 탐지하고, 테이블의 내용을 바탕으로 핵심 카테고리와 코사인 유사도 및 Rouge Score를 비교하여 핵심 영역을 식별한다.

몇 가지의 한계점이 존재하는데, 다음과 같다. 첫 번째는 핵심 카테고리 별로 데이터를 수집하는데 많은 시간이 소요된다. 비정형 문서이기 때문에 파일들을 하나씩 열어보고, 핵심 카테고리에 해당하는 테이블이 존재하는지 확인하며 라벨링 해야하는 번거로움이 존재한다.

두 번째는 복잡하거나 구분선이 없는 테이블일 경우, YOLOv8 detection에서 미탐이 발생할 수 있다. 연구에 사용한 데이터에는 해당하는 데이터가 거의 존재하지 않았기 때문에 추가로 학습시키지는 않았다. 필요한 경우 추가 학습을 진행한다면 한계점을 개선할 수 있을 것으로 보인다.

세 번째는 코사인 유사도가 높게 나와도 핵심 키워드가 적은 경우 미탐이 발생할 수 있다. Rouge Score를 적용할 때, threshold보다 값이 낮다면 식별되지 않기 때문이다. 핵심 키워드를 행 제목 혹은 열 제목으로 선정하였기 때문에, 핵심 카테고리 별로 라벨링하는 테이블들은 어느 정도 공통된 형식을 가지고 있는 것이 좋다.

향후 연구로는 핵심 영역을 표로만 제한하지 않고 문서 영역 식별 모델을 도입하여 표 이외의 문서 영역에서도 핵심 영역을 식별할 수 있도록 보완할 필요가 있다. 또한, rouge score를 사용하지 않고도 테이블에서 key와 value를 매핑 시켜주는 모델을 도입하여 새로운 방법으로 테이블 안에서의 핵심 정보를 식별하는 방향도 고려해 볼 수 있을 것이다.

참고문헌

[1] 김효종, "기밀 문서 파일 유출 방지를 위한 FCLPS에 관한 연구," 국내석사학위논문 동명대학교 대학원, 2021.

[2] 홍용기, 고기혁, 양희동, and 류승환, "프라이버시 보호 데이터 배포: 정형 및 비정형 데이터 비식별화 기술 동향," 정보과학회논문지, Vol. 50, No. 11, pp. 1008-1029, 2023.

[3] 양병모 and 양오석, "Word2Vec 모델을 이용한 ESG 점수 도출에 관한 연구: 비정형 문서간 유사도 분석을 활용한 텍스트 계량화 방법론 제안," 경영연구, 37, pp. 57-72, 2022.

[4] 김원준. "특허와 논문정보를 활용한OCR기술발전 동향예측에 관한연구." 국내박사학위논문 한국기술교육대학교 일반대학원, 2023. 충청남도

[5] dreamgonfly. "쉽게 씌어진 word2vec" https://dreamgonfly.github.io/blog/word2vec-explained/#%EB%8B%A8%EC%96%B4-%EC%9E%84%EB%B2%A0%EB%94%A9word-embedding-%EB%A7%9B%EB%B3%B4%EA%B8%B0. Accessed: 2024-01-15.

[6] 서혜선 (2020). LDA와 Word2vec 방법론을 이용한 의정부시 SNS 데이터의 토픽 모델링 및 시각화, Journal of The Korean Data Analysis Society, 22(6), 2391-2403.

[7] 강형석 and 양장훈, "Word2vec 및 fastText 임베딩 모델의 성능 비교," 디지털콘텐츠학회논문지, Vol. 21, No. 7, pp. 1335-1343, 2020.

[8] 전민규 and 김남규, "텍스트 요약 품질 향상을 위한 의미적 사전학습 방법론," 스마트미디어저널, Vol. 12, No. 5, 17-27, 2023.