문장 의도 분류와 개체명 인식을 활용한 개인정보 검출 및 비식별화 시스템

서동국*, 김건우**, 김재영***, 이동호*
*한양대학교 컴퓨터공학과
**청운대학교 AI운영학과
***한양대학교 ERICA 소프트웨어학부
*{sdk6789, dhlee72}@hanyang.ac.kr, **gwkim@chungwoon.ac.kr, ***kjyl95718@hanyang.ac.kr

Personal Information Detection and De-identification System using Sentence Intent Classification and Named Entity Recognition

Dong-Kuk Seo*, Gun-Woo Kim**, Jae-Young Kim***, Dong-Ho Lee*
*Dept. of Computer Science and Engineering, Hanyang University
**Dept. of Artificial Intelligence Operations, Chungwoon University
***Dept. of Software, Hanyang University ERICA

요 약

최근 개인정보가 포함된 비정형 텍스트 문서들이 유출되거나 무분별하게 공개됨으로써 정보의 주체는 물론 기업들까지 피해를 받고 있다. 데이터를 공개 및 활용하기 위해 개인정보 검출 및 비식별화 과정이 필수적이지만 정형 데이터와는 달리 비정형 데이터의 경우 해당 과정을 자동으로 처리하는 데 한계가 있다. 이를 위해 딥러닝 모델들을 사용하여 자동화하려는 연구들이 있었지만 문장 내 단어의 모호성에 대한 고려 없이 단어 개체명 정보에만 의존하여 개인정보를 검출하는 형태로 진행되었다. 따라서 문장 내 단어들 중 식별 대상인 단어들도 비식별화 되어 데이터에 대한 유용성을 저해할 수 있다는 문제점을 남겼다. 본 논문에서는 문장의 의도 정보를 단어의 개체명 학습 과정에 부가적인 정보로 활용하는 개인정보 검출 모델과 개인정보 데이터의 유용성을 고려한 비식별화 기법을 제안한다.

1. 서론

기업에서 다루는 문서들에는 개인의 신원을 파악하는 용도로 개인정보들이 다수 포함되는 것이 보통이다. 하지만 최근 이러한 개인정보들이 유출되어 유출된 정보의 주체는 물론 기업 또한 피해를 받는 사례들이 증가하고 있다.

2016년 발행된 '개인정보 비식별 조치 가이드라인'은 개인정보를 식별자와 속성자로 정의하고 있다.[1] 식별자는 성명과 같이 단일 정보만으로 개인을 특정할 수 있는 정보이며 속성자는 다른 정보와 쉽게 결합하여 개인을 특정할 수 있는 정보를 말한다. 데이터 3법의 개정으로 인해 기업은 해당 지침에 따라 비식별 조치를 선행한 후 데이터를 공개해야 하지만 해당 정보를 사람이 직접 선별하는 데는 많은 비용과 어려움이 따른다.

개인정보 비식별화 과정을 자동으로 처리하기 위한 기존 연구들은 대부분 레코드 단위의 정형 데이터를 기반으로 연구되었다. 하지만 기업이 다루는 문서들 중에는 자기소개서, 이력서 등과 같이 비정형 텍스트 문서들도 적지 않기 때문에 최근 가시적인 성능을 보이는 딥러닝 모델을 활용하여 문서 내부의 개인정보를 자동으로 비식별화하려는 연구가 진행되었다. 그러나 단순히 문장 내 단어에 대한 개체명 인식 결과에 의존하여 비식별화를 진행하였기 때문에 개인정보가 아닌 정보까지도 비식별 처리되어 데이터의 유용성을 저하시킬 수 있다는 문제점을 남겼다.

본 논문에서는 End-to-End의 단일 모델 형태로써 문장의 의도 정보를 분류하고 이를 문장 내 단어에 대한 개체명 학습에 활용하는 개인정보 검출 모델을 제안한다. 또한 검출된 개인정보에 대한 비식별 조치를 위해 데이터의 유용성을 보호할 수 있는 비식별화 기법을 제안한다.

2. 관련 연구

개인정보를 비식별화하려는 노력은 비교적 오래전부터 진행되어 왔다. k-익명성, i-다양성, t-근접성 모델은 개인정보 노출에 대한 정량적인 위험성을 규정하여 개인정보 유출을 최소화하고자 하였다.[2] 하지만 해당 연구들은 단순히 데이터의 주체를 식별할 수 없도록 변형을 가하는 방식으로써 데이터의 2차

적 활용이 어렵다는 점과 비정형 텍스트 데이터에는 적용할 수 없다는 문제점이 있었다.

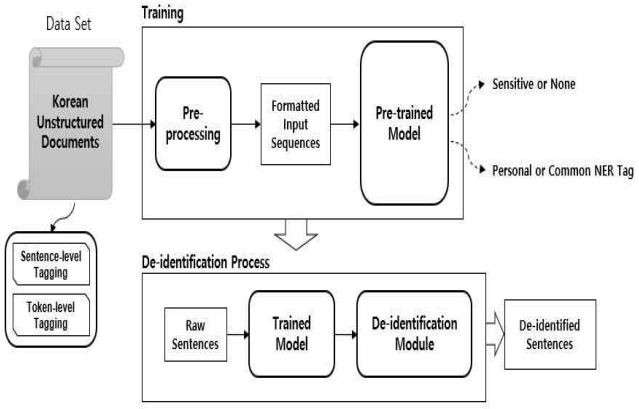
따라서 비정형 텍스트 데이터 비식별화에 초점을 둔 연구가 진행되었다.[3] 해당 연구는 의료 지식베이스 임베딩 벡터 및 사전 지식들을 추가 자질로서 활용하여 개인정보 개체들을 자동으로 검출하고자 하였다. 하지만 해당 연구의 추가 자질들은 다른 도메인에 적용할 수 없다는 점과 단순 개체명 인식 결과에만 의존하여 비식별화를 진행하였기 때문에 비식별 대상이 아닌 개체들까지도 모두 비식별화됨으로써 데이터의 유용성을 저하시킨다는 문제가 존재하였다.

[4]는 비정형 문서 내 존재하는 테이블의 단어들을 문장으로 생성한 후 개체명 인식 학습을 통해 개인정보를 검출하였다. 하지만 해당 연구는 제한된 수의 문서 형식 및 테이블에서 사용된 단어들에 대해서만 학습하여 비식별화를 진행하였기 때문에 새로운 단어에 대해서는 개인정보 여부를 판별하지 못하는 OOV(Out-of-Vocabulary)문제가 발생하였다.

본 논문에서는 비정형 텍스트 데이터에서 등장하는 다양한 단어들에 대한 학습 정보를 바탕으로 개인정보를 검출하기 위해 한국어 사전 학습 언어 모델들을 활용하여 단일 모델에서 문장의 의도 분류와 개인정보 검출을 동시에 진행한다.

3. 개인정보 비식별화 시스템

3.1 시스템 전체 구조



(그림 1) 제안 시스템 전체 구조도

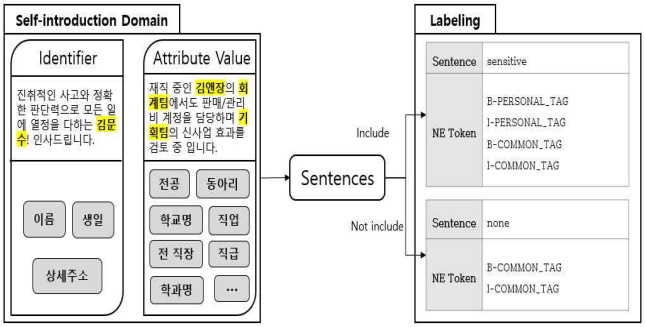
그림 1은 본 논문에서 제안하는 시스템의 전체 구조도이다. 모델의 입력으로 한국어 기반의 비정형 문서가 사용되며 문서 내 문장들은 문장 및 토큰 단위로 태깅된다. 각각의 문장 및 토큰들이 모델에 입력되면 문장의 의도와 각 토큰의 개체명을 동시에 학습한다. 그 후 학습된 모델의 추론 결과를 바탕으로 각 토큰에 대한 비식별화가 진행된다.

3.2 데이터 셋

<표 1> 개인정보 개체명 태그 리스트

Tag	Category	Definition
PER	사람	사람 이름
ORG	학교	학교 이름
	기업	기업 이름
EDU	자격증	소지 자격증
	전공	졸업, 재학 전공 학과
AFF	업무 부서	이전, 지원 업무 부서
	동아리	이전, 현재 소속 동아리
POS	직업	이전, 현재 직업
	직위	부서 내 직위
LOC	나라	나라 이름(유학 정보)
	지역	주소지
DUR	기간	특정 기간

본 논문에서 검출하고자 하는 문서 내 개인정보는 표 1의 7개 개체들이며 [1]에서 정의하는 식별자, 속성자의 특성 범주를 기반으로 해당 도메인에 부합하는 식별자, 속성자를 별도로 정의한다.



(그림 2) 개인정보 정의 및 태깅 방법

그림 2는 자기소개서 문서를 입력으로 사용할 때 해당 도메인에 정의된 개인정보(식별자, 속성자)와 태깅 방법에 대한 그림이다. 문장 단위 태깅의 경우 해당 문장이 개인에게 민감한 문장인지 아닌지에 대한 태깅으로서 문장 안에 개인정보를 하나라도 포함할 경우 'Sensitive', 하나도 포함하지 않을 경우 'None'으로 태깅한다.

<표 2> 개인정보 개체 태깅 및 제외 문장 예시

Case 1	"코리아텍에서 보안팀 주임**(B-PERSONAL_POS)**으로 일하던 중 다른 부서 차장**(B-COMMON_POS)**님께서 다급히 연락이 왔던 적이 있었습니다."
Case 2	"따라서 **디자인팀**에서 먼저 웹사이트 메인 배너 디자인 교체 필요성을 주장했습니다."

토큰 단위 태깅의 경우 개체명 인식 분야에서 사용되는 정답 인코딩 방식인 B-I-O 태깅 방법을 적용한다. B(Begin)는 개체명의 시작 토큰일 경우,

I(Inside)는 연장되는 개체일 경우, O(Outside)는 학습하고자 하는 개체명 토큰이 아닐 경우 부착된다. 만약 한 문장 내에 개인정보 개체와 아닌 개체가 동시에 출현한 경우 두 개체를 구분하여 학습시키기 위해 표 2의 첫 번째 예시와 같이 별도의 태그 구분자를 사용한다. 개인정보 개체는 'PERSONAL', 타인 혹은 공통적인 정보의 개체는 'COMMON' 구분자를 부여한다.

또한 표 2의 두 번째 예시와 같이 문장에서 표 1에 해당하는 개체가 등장하더라도 해당 개체를 개인정보라고 확정할 수 없는 경우 해당 문장은 데이터셋에서 제외한다.

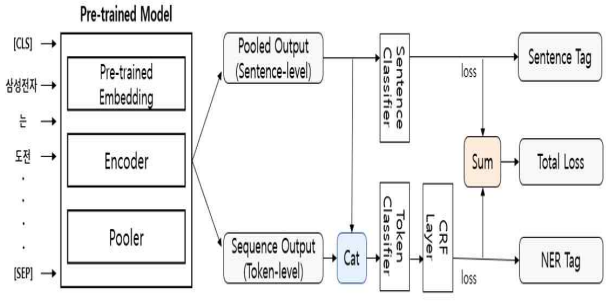
3.3 사전 학습 모델

최근 NLP 분야 관련 기술의 비약적인 발전은 사전 학습된 언어 모델을 통해 이루어졌다. 사전 학습 언어 모델은 고성능의 하드웨어를 이용하여 대량의 텍스트 데이터를 학습시켜 놓은 것을 말하며 다양한 응용 태스크에서 이미 우수한 성능을 입증하고 있다.

본 논문에서는 보다 정확한 한국어 개인정보 검출을 위해 대량의 한국어 텍스트 데이터로 사전 학습된 BERT, KoBERT, DistilKoBERT 사전 학습 모델을 활용한다.[5-7]

3.4 개인정보 검출 모델

개체명 인식은 문장의 각 토큰이 어떤 개체에 해당하는지 판단하는 작업으로 문맥에 대한 분석이 중요하게 작용한다. 기존 개체명 인식 모델은 정답 태그가 태깅된 개체들의 주변 토큰들을 하나의 문맥으로 삼아 개체명을 학습한다. 하지만 단순히 해당 개체의 클래스를 분류하는 문제가 아닌 동일 범주 클래스 개체들에 대한 비식별 여부를 판단하기 위해서는 보다 높은 단계의 문맥 정보가 필요하다.



(그림 3) 제안 모델 구조

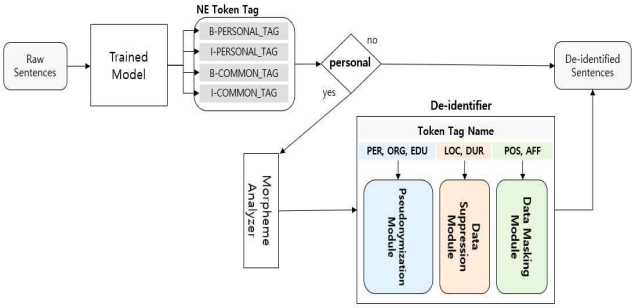
그림 3은 본 논문에서 제안하는 개인정보 검출 모델의 구조이다. 제안하는 모델은 개인정보와 개인정보가 아닌 개체들이 한 문장 안에 공존할 때 이를 보다 정확하게 분류 및 검출하기 위해 문장 단위의 예측 정보를 추가적인 자질로서 활용한다. 즉 해당 문장이 개인에게 민감 혹은 민감하지 않다고 판단된 문맥 자질 요소들을 개인정보 개체 검출에 부가적인 정보로서 활용한다.

모델의 산출 결과 중 'pooled output'은 문장 단위의 은닉 상태 결과 값이며 해당 결과 값은 시퀀스 전반의 정보를 함축하고 있는 '[CLS]' 토큰을 이용하여 산출된다. 또 다른 산출 결과인 'sequence output'은 시퀀스의 각 토큰들에 대한 은닉 상태 결과 값이다. 산출된 두 결과 값이 결합된 정보를 기반으로 정답 시퀀스 사이의 의존성을 학습하는 CRF층을 거쳐 각 토큰의 개체명을 학습한다.

3.5 개인정보 비식별화

개인정보에 대한 비식별화는 제안 모델의 개체명 추론 결과를 토대로 이루어진다. 단순 마스킹 형식의 비식별화를 통해 소실되는 정보들을 보존하고자 추론 결과에 따른 별도의 비식별 모듈들을 이용한다.

또한 문장 단위의 잘못된 추론 결과로 인해 해당 문장 안에 존재하는 개인정보를 비식별 처리하지 못하는 문제를 염려하여 문장 추론 결과는 비식별화 과정에 활용하지 않는다.



(그림 4) 개인정보 비식별화 과정

그림 4는 검출된 개인정보에 대한 비식별화 과정이다. 추론된 개체명이 개인정보일 경우 개체의 경계를 명확하게 구분 짓기 위해 한국어 형태소 분석기를 활용한다. 형태소 분석기에서 정의하는 조사, 어미로 분류된 형태소 태그 사전을 별도로 정의하고 스캔 매칭을 수행하여 토큰에서 조사, 어미를 분리한다.

<표 3> 개인정보 비식별화 예시

Before	"안녕하십니까. **한양대학교(ORG) 김석사(PER)**입니다. 저는 **경기도 안산시 상록구(LOC)**에 거주하고 있으며 **2003년(DUR)**부터 **동원F&B(ORG) 영업부(AFF) 대리(POS)**로 근무했던 경력이 있습니다."
After	"안녕하십니까. **한국대학교 홍길동**입니다. 저는 **경기도**에 거주하고 있으며 **2000년대 초반**부터 **A회사 OOO OO**로 근무했던 경력이 있습니다."

그 후 조사, 어미가 분리된 개체 토큰은 추론 결

과에 따라 정의된 가명화, 범주화, 마스킹 모듈을 거쳐 표 3의 예시와 같이 비식별 처리된다.

4. 실험 및 결과

4.1 실험 환경

제안하는 시스템의 실험을 위해 '인크루트', '사람인'과 같은 구인구직 사이트에 공개된 자기소개서 문서들의 10,025개 문장들을 활용하였다. 이 중 80%인 8,020개의 문장은 학습 데이터 셋으로 사용되었으며 20%인 2,005개의 문장은 테스트 데이터 셋으로 사용되었다. 또한 학습 데이터 셋의 10%인 802개의 문장은 검증 데이터 셋으로 사용하였고 모든 모델의 패딩, 에포크, 배치 사이즈는 각각 80, 20, 32로 설정하였다.

4.2 실험 결과

제안하는 시스템의 검증을 위해 문장 의도 분류 정보를 개인정보 검출에 추가적인 정보로 사용하지 않는 모델과 본 논문에서 제안하는 모델의 정확도를 비교하였다. 각 모델의 비교 대상은 문장 의도 분류 정확도와 개인정보 개체명 인식 F1 점수, 그리고 비식별화 정확도를 나타내는 외부 문장들에 대한 모델 추론 정확도이다.

<표 4> 모델 별 성능 비교

	Sentence Intent Accuracy	NER F1-Score	De-identification Accuracy
BERT	0.82	0.7763	78.0%
DistilKoBERT	0.81	0.7749	84.6%
KoBERT	0.84	0.8135	86.9%
Proposed Model (BERT)	0.82	0.7967	79.6%
Proposed Model (DistilKoBERT)	0.81	0.8072	85.0%
Proposed Model (KoBERT)	0.84	0.8449	88.8%

표 4는 각 모델의 성능을 비교한 결과이다. 문장 의도 분류 정확도에서는 한국어 학습 비중이 가장 큰 KoBERT가 가장 높은 성능을 보였다. 비교 모델들 중 한국어로만 학습된 DistilKoBERT의 성능이 BERT보다 높을 것이라 예상했으나 비교적 적은 한국어 코퍼스 및 모델 파라미터 학습으로 인해 문장 분류 정확도 및 개체명 인식 F1 점수 모두 더 떨어지는 결과를 보였다. 하지만 제안 모델의 F1 점수에서는 DistilKoBERT의 성능이 BERT보다 더 높게 측정되었으며 제안 모델들 중 KoBERT의 점수가 0.8449로 가장 높은 결과를 보이면서 제안 모델들의 F1 점수가 비교 모델들 대비 적게는 0.02, 많게는 0.032 가량 향상된 것을 확인하였다.

비식별화 정확도는 별도로 구축한 100개의 문장에서 표 1 태그들에 부합하는 260개의 타겟 토큰들을 선정하여 평가하였다. 타겟 토큰의 태그를 모델이 추론하지 못하거나 정답과 다르게 추론한 경우 틀린 것으로 간주하였고 타겟 토큰이 아닌 토큰이 잘못 추론된 경우는 정확도 계산에서 배제하였다.

비식별화 실험 결과 본 논문에서 제안한 모델들의 성능이 비교 모델들 보다 향상되었음을 확인할 수 있었으며 그중 KoBERT를 사전 학습 모델로 사용한 모델의 정확도가 88.8%를 달성하며 가장 높은 정확도를 보였다.

5. 결론 및 향후 연구

본 논문에서는 한국어 사전 학습 언어 모델을 기반으로 문장의 의도 정보를 개인정보 검출에 부가적인 정보로 활용하는 모델과 데이터의 유용성을 고려한 비식별화 기법을 제안하였다. 또한 실험을 통해 본 논문에서 제안한 시스템의 우수성을 검증하였다.

향후 연구로는 지식 그래프와 같은 추가적인 정보들을 활용하여 제안된 모델을 고도화시키는 연구와 비식별화 모듈 내 정의되는 사전 지식들을 자동으로 구축하는 연구에 대하여 진행할 예정이다.

이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2020001343,인공지능융합연구센터(한양대학교 ERICA))

이 논문은 과학기술정보통신부의 소프트웨어중심대학 지원사업(2018-0-00192)의 지원을 받아 수행하였음

참고문헌

[1] 개인정보 비식별 조치 가이드라인(2016.06, 국무조정실, 행정자치부, 방송통신위원회, 금융위원회, 미래창조과학부, 보건복지부)

[2] P. Twinkle, A. Kiran "A Study on k-anonymity, I-diversity and t-clossness Techniques of Privacy Preservation Data Publishing", International Journal for Innovative Research in Science & Technology, Vol. 6, Issue 6, 2019.

[3] X. Yang et al, "A Study of Deep Learning Methods for De-identification of Clinical Notes at Cross Institute Settings", BMC Medical Informatics and Decision Making, 19, 232, 2019.

[4] J. Park et al, "Sensitive Data Identification in Structured Data through GenNER Model based on Text Generation and NER", CNIOT, pp 36-40, 2020.

[5] D. Jacob et al, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv:1810.04805, 2018.

[6] SKTBrain, "Korean BERT pre-trained cased", Github repogitory, Github, https://github.com/SKTBrain/KoBERT, 2019.

[7] Park. Jangwon, "DistilKoBERT: Distillation of KoBERT", Github, https://github.com/monologg/DistilKoBERT, 2019.