개인정보 탐지를 위한 특화 개체명 주석 데이터셋 구축 및 분류 실험

강혜린°, 비립, 강예지, 박서윤, 조예슬◇, 성현민◇, 장성순◇, 김한샘†
연세대학교 언어정보연구원†, 월드버텍◇

{hyerink°, feili0820, yjkang5009, seoyoon.park, khss†}@yonsei.ac.kr,
{yesl-c, sung8504, madflame}◇@vtex.co.kr

A Study on the Construction of Specialized NER Dataset for Personal Information Detection

Hyerin Kang°, Li Fei, Yejee kang, Seoyoon Park, Yeseul Cho◇, Hyeonmin Seong◇, Sungsoon Jang◇, Hansaem Kim†

Institute of Language and Information Studies, Yonsei University†
WorldVertex◇

요 약

개인정보에 대한 경각심 및 중요성 증대에 따라 텍스트 내 개인정보를 탐지하는 태스크가 주목받고 있다. 본 연구에서는 개인정보 탐지 및 비식별화를 위한 개인정보 특화 개체명 태그셋 7개를 고안하는 한편 이를 바탕으로 비식별화된 원천 데이터에 가상의 데이터를 대치하고 개체명을 주석함으로써 개인정보 특화 개체명 데이터셋을 구축하였다. 개인정보 분류 실험에는 KR-ELECTRA를 사용하였으며, 실험 결과 일반 개체명 및 정규식 바탕의 규칙 기반 개인정보 탐지 성능과 비교하여 특화 개체명을 활용한 딥러닝 기반의 개인정보 탐지가 더 높은 성능을 보임을 확인하였다.

주제어: 개인정보, 개체명, 개인정보 특화 개체명, 개인정보 탐지

1. 서론

현재 민감 정보를 포함하는 많은 문서들이 유출되고 있으며, 이에 따른 피해는 개인뿐만 아니라 기업과 공공기관에도 막대하기 때문에 민감 정보를 판별하고 이에 대한 마스킹 처리가 필수적으로 이루어져야 한다.

따라서 이러한 개인정보 데이터를 활용할 때에 중요한 것이 대량의 데이터에 포함된 개인정보를 탐지하고 비식별화하는 것이다. 개인정보 비식별화란 데이터 내 특정 개인을 식별할 수 있는 정보가 있을 경우에 개인정보의 일부 혹은 전부를 삭제하거나 속성을 대체하여 처리하거나 다른 정보와 결합하였을 때에도 특정 개인을 식별하는 것을 방지하는 것이다. 이를 위해 반드시 개인정보의 올바른 탐지가 선행되어야 한다. 정확한 개인정보 탐지가 우선되어야 탐지 결과를 바탕으로 비식별화를 진행할 수 있기 때문이다.

데이터 내에 포함된 개인정보를 탐지하고 처리하기 위해 개체명이 주로 활용된다. 개체명 인식 기술을 활용하여 적절한 개체를 자동으로 인식하고 인식 결과를 토대로 개인정보와 연결 짓는 것이다. 그러나 이는 범용적인 개체명 태그셋을 활용하여 개인정보를 탐지하는 것으로 일반 개체명 태그셋의 유형과 정의로는 포함되지 못하는 개인정보의 유형이 존재한다. 이는 개인정보 탐지 성능을 저하시키는 요인으로 작용할 수 있다. 때문에 개인정보 탐지를 위해서는 기존의 일반적인 개체명을 개인정보적 관점에서 특화시킨 새로운 개체명 체계가 필요하다.

따라서 본 연구는 개인정보 탐지 및 비식별화를 위한 개인정보 특화 개체명 주석 체계 고안의 기초 연구로서 일부 개인정보 특화 개체명 주석을 기반으로 데이터셋을 구축한 후, 실험을 진행하였다. 실험은 일반 개체명 태그셋 데이터셋과 개인정보 태그셋 데이터셋을 대상으로 하여 실험을 진행하였으며 분류 성능을 비교 및 검증하고 개인정보 탐지를 위한 개인정보 특화 개체명의 효율성을 정성적 평가를 통해 검토하였다.

2. 관련 연구

2.1 개인정보 탐지 및 비식별화

문서 내에서 개인정보 탐지를 시도한 연구에는 [1]과 [2]를 들 수 있다. [1]에서는 NER 을 이용하여 법률 문서 내에서 개체를 식별하고 민감한 정보를 자동으로 인식하는 실험을 진행하였으며, 이를 위해 generic NER 을 이용하는 한편 개체가 잘 식별되었는지 평가하기 위해 해당 단어를 적절한 컨텍스트 윈도우(context window)에 포함시켜 같은 도메인의 택소노미(domain

taxonomy)에서 비슷한 단어와의 유사도를 계산하였다. 이를 통해 카테고리의 라벨을 정교화 하였으며, 개체가 잘 식별되었는지 평가하기 위해 적절한 컨텍스트 윈도우에 해당 단어를 포함시켜 도메인 택소노미에서의 유사한 단어가 도출되는지를 평가하였다. 문서에 포함된 비정형 데이터는 물론 정형 데이터 내 개인정보 탐지를 위한 국내 연구에는 [2]를 들 수 있다. [2]에서는 딥러닝 발달에 따라 문맥 정보를 가진 비정형 데이터의 민감정보 식별 성능이 향상되었지만, 문맥 정보가 결여된 정형 데이터(예: 표)에서는 낮은 탐지 성능을 보이는 것을 지적하며 표 정보 내 개체명을 식별함으로써 민감정보를 판별하는 실험을 진행하였다. 이를 위한 데이터셋으로 서울시 정부 문서 내 100 개 가량의 표를 사용하였고, 양방향 LSTM 모델과 표에 등장하는 단어를 활용하여 문서에 등장할 법한 문장을 의도적으로 생성함으로써 문맥 정보를 생성하였다. 생성 이후 개체명을 식별하여 해당 단어가 표 안에서 민감정보인지 아닌지를 판별한 결과 문장을 통한 문맥 생성 후 표에 나타나는 민감 정보 개체명 탐지 성능이 높아지는 것을 확인하였다.

데이터 사용범위에 따라 개인정보 비식별화 정도 및 처리 방법을 연구한 논문으로는 [3]이 있다. [3]에서는 데이터 사용범위를 기능적 활용단계와 연계형 활용단계, 그리고 보장형 활용단계로 나눈다. 기능적 활용단계는 단순 연구 데이터로써 데이터를 사용하는 경우이며, 비식별화 역시 단순 연구 데이터에 준하여 빠른 처리 속도와 재활용 가능성을 우선으로 한다. 그러나 연계형, 보장형 활용단계부터는 한 기관에서부터 여러 기관까지의 기관들이 단위가 되기에 추가 정보에 의한 개인정보 유추 가능성을 차단하는 한편, 비식별화 후 추가 데이터와의 결합으로 개인정보 유추가 불가능하도록 하는 목적을 띈다. 본 연구는 데이터 내 등장한 개인정보 비식별화에 대한 초기 연구로써 [3] 연구의 기능적 활용단계에 속한다. 그러나 궁극적인 목표는 데이터 내 개인정보 비식별화 및 탐지를 통한 데이터 비식별화와 이의 상업적 활용도 제고로써 향후 연구에 있어 지속적이고 심도 있는 비식별화를 목표로 한다.

[4]에서는 단순 문장 내 단어에 대한 개체명 인식 결과만을 가지고 비식별화를 진행하게 될 때에 개인정보 외의 정보도 비식별 처리되어 데이터 유용성이 제한된다는 한계점을 지적하였다. 따라서 본 연구에서는 문장의 의도 정보 분류 결과를 개체명 학습에 활용하는 개인정보 검출 모델을 제안하였다. 제안한 모델을 실험을 통해 정확도를 검증하였다. 실험 결과 문장 의도 분류를 학습의 추가 정보로 활용한 모델의 경우 그렇지 않은 모델들 보다 성능이 향상되는 것을 확인할 수 있었다.

2.2 개인정보 관련 개체명 연구

[5]에서는 딥러닝을 활용한 개인정보 비식별화를 위하여 개인정보 관련 개체명을 재정의하였다. 또한 재정의한 개체명을 토대로 한국어 위키피디아 본문을 활용하여 비식별화 데이터 생성 방법을 제안하고 이 방법에 따라 총 3,000 문장의 데이터를 생성하였다. 비식별화를 위하여 재정의한 개체명은 총 12 개의 유형으로 구성되어 있으며, 'PER'를 '사람 이름, 별명, 캐릭터, 가수/그룹명' 과 같이 개인을 식별할 수 있는 정보를 포함하고 있는 경우에 해당하는 것을 포함시켜 재정의하였다. 언어 모델의 결과에서 재정의한 개체명 유형 중 대응되지 않는 경우에 대해서는 개체명 태깅을 삭제한 후 반자동 시스템에서 치환되도록 하였다. 모델에서 개체명으로 인식되지 않는 혈액형, 병명, 전공, 학력 등의 정보는 규칙을 부여하여 개체명 태그를 부착하였다.

[6]에서는 기존 개인정보 탐지를 위해 이루어진 패턴 및 형태소 분석의 한계점을 지적하며 기계학습 알고리즘을 이용한 개인정보 식별율 향상 기법을 제안하였다. 이를 실험하고자 개인을 식별할 수 있는 정보들을 개인정보 비식별화 대상으로 삼고 개인정보 태그셋을 구축하였다. 이를 토대로 구축한 학습 데이터셋을 통해 기계에게 다양한 패턴 및 유형을 학습시킨 후 기계학습 알고리즘에 변화를 주어 개인정보를 재추출하도록 하는 반복적 실험을 거쳐 개인정보 식별율의 향상 정도를 평가하였다. 그 결과 패턴 위주의 개인정보 식별 방식보다 기계학습을 통한 지속적인 재학습을 사용할 때 식별율이 더욱 향상된 것을 확인하였다.

[7]에서는 의료분야(clinic) 데이터셋 내 등장하는 개인정보 관련 개체명을 탐지하고자 teacher-student model 을 사용하여 실험을 진행하였다. 데이터셋으로는 프랑스어 의료분야 데이터셋들을 사용하였으며, 각 데이터셋에서 쓰이는 개체명 태그들을 통합하고자 MERLOT 데이터셋의 6 개 개체명을 기준으로 하여 개체명을 통합하였다. 실험 결과 student 모델은 70.6%의 F1-score 를 달성하여 개인정보 개체명 데이터셋의 필요성을 환기하는 한편 개인정보 개체명 자동 주석의 가능성을 보였다.

[8]에서는 사용자 스토리(user story) 데이터셋에 개체명 주석을 진행하여 사용자 스토리 내 프라이버시(privacy) 정보에 속하는 개인 데이터 속성(personal data attributes), 데이터 주체(data subject) 등과 관련된 개체명 탐지 정도를 실험하였다. 이를 위해 최신의 개체명 모델과 문맥 임베딩(contextual embedding)을 결합한 모델을 사용하였으며 세 가지 개체명 탐지에 대해 유효한 결과를 보였다.

3. 개인정보 특화 개체명 주석 데이터셋 구축

3.1 데이터 수집

원천 데이터(raw data)로는 AI Hub 의 '용도별 목적대화' 말뭉치를 선정하였다. '용도별 목적대화' 말뭉치는 여러 분야의 고객 상담 대화로 구성된 데이터셋으로, 다양한 플랫폼의 고객 문의 및 주문, 예약 등을 수행하는 목적별 대화들이 수집되어 있다. 때문에 고객이나 상담원의 이름, 소속, 주소, 전화번호 등이 비식별화 되어있으며, 본 연구의 목표인 개인정보 특화

개체명과 개인정보 식별 여부를 실험할 수 있는 적합한 데이터이기에 원천 데이터로 선택하였다. '용도별 목적대화' 말뭉치의 비식별화 목록은 표 1.과 같다.

표 1. '용도별 목적대화' 말뭉치 비식별화 목록

비식별 태그	내용
#@이름#	사람 성명
#@소속#	직장, 단체 등 소속
#@주소#	건물, 시설 등의 주소 일체
#@전번#	휴대전화 번호, 일반 전화 번호
#@URL#	웹 페이지 주소
#@계정#	웹 사이트 ID, 이메일 주소
#@번호#	자동차 번호, 호실 등 번호 일체
#@금융#	신용카드 번호, 계좌 번호

본 연구에서는 모델 학습과 실험을 위해 AI Hub 에 공개된 해당 용도별 목적대화의 20 여개 주제 중 약 130 만 어절(222,157 문장)을 포함하는 12 개 주제에 대한 말뭉치를 선택하여 실험을 위한 원천 데이터셋으로 사용하였다.

3.2 개체명 데이터셋 구축

실험을 위해 구축한 데이터셋은 '일반 개체명 데이터셋'과 '개인정보 특화 개체명 데이터셋'이다. 이를 위해 3.1 에서 수집한 AI Hub 용도별 목적 대화 말뭉치의 비식별화 항목에 대해 가상 데이터를 생성하는 faker 패키지(이름, 이메일, URL)로 대치하거나 임의의 숫자 조합으로 생성한 데이터(휴대전화 번호, 일반전화/FAX)로 대치하였다. 전화번호는 휴대전화 번호와 일반전화/FAX 번호에 따라 개인 식별 위험도에 차이가 있을 수 있으므로, 개체명을 분리하였다. 한편 주소, 직장명에 대해서는 기존에 공시된 주소명(주소기반산업지원서비스), 기업 명단(상장법인목록, 벤처기업명단) 데이터를 활용하여 항목을 대치함으로써 개인정보를 갖춘 데이터를 생성하였다. 이때 비식별 태그 '#@번호#'에 속하는 차량번호, 그리고 '#@금융#'에 속하는 카드번호, 계좌번호는 말뭉치에 포함된 데이터 건수가 희소하여 대치 대상에서 제외하였다.

해당 데이터에 대해 '일반 개체명' 태그셋과 본 연구에서 고안한 '개인정보 특화 개체명' 태그셋으로 개체명 태깅을 진행함으로써 최종적으로 연구에 사용할 개체명 데이터셋을 구축하였다. '일반 개체명 데이터셋' 구축 시에는 기존 국립 국어원에서 공개한 '개체명 말뭉치'에서 사용된 개체명 태그셋 15 개를 사용하였다. 한편 '개인정보 특화 개체명'은 국립국어원의 15 개 개체명 분류에 개인 정보가 포함되는지 여부를 분석하여 고안한 7 개 개체명으로써, 목록은 표 2.와 같다.

표 2. 비식별 태그와 개인정보 특화 태그셋

비식별 태그	개인정보 특화 태그셋
#@이름#	PS_NAME
#@소속#	OG_WORKPLACE
#@전번#	QT_PHONE QT_MOBILE
#@URL#	TMI_SITE
#@주소#	LC_ADDRESS
#@계정#	TMI_EMAIL

실험을 위해 학습용 데이터셋과 평가용 데이터셋은 데이터셋 특성상 단순히 대화셋을 9:1 로 랜덤 분할할 경우 이메일, URL 과 같이 빈도가 적은 개체명에서 편향이 생길 위험이 있어, 각각의 개체명 항목들의 밸런스를 유지하면서 대화셋 전체에 대한 비율 또한 90%, 10%로 나누어질 수 있도록 학습용 데이터셋과 평가용 데이터셋을 수동으로 분리하였다. 개인정보 특화 개체명과 일반 개체명들의 학습 및 평가용 데이터셋 분포는 표 3-1.과 표 3-2.와 같다.

표 3-1. 개인정보 특화 개체명 태그셋 분포

개인정보	개체명 태그	학습용 데이터셋	평가용 데이터셋	전체
이름	PS_NAME	22,284	2,706	24,990
이메일	TMI_EMAIL	63	10	73
URL	TMI_SITE	333	26	359
주소	LC_ADDRESS	9,751	1,198	10,949
직장명	OG_WORKPLACE	25,386	3,120	28,506
핸드폰번호	QT_MOBILE	429	47	476
일반전화/FAX	QT_PHONE	2,559	288	2,847

표 3-2. 일반 개체명 태그셋 분포

개인정보	개체명 태그	학습용 데이터셋	평가용 데이터셋	전체
이름	PS	23,340	2,641	25,981
이메일, URL	TM	7,076	844	7,920
주소	LC	15,614	1,831	17,445

직장명	OG	25,965	2,799	28,764
핸드폰번호,일반전화/FAX	QT	19,985	2,262	22,247

4. 실험

실험은 '일반 개체명 태그셋'과 '개인정보 특화 개체명 태그셋'을 사용한 각각의 데이터셋에 대해 진행하였으며, '개인정보 특화 개체명 태그셋' 실험의 경우, 규칙 기반(rule-based)과 딥러닝 기반으로도 실험을 진행한 후 결과를 비교하였다.

4.1 실험 설계

본 논문에서는 서울대학교 컴퓨터언어학 연구실에서 공개한 한국어 특화 ELECTRA 모델인 KR-ELECTRA 모델을 Pre-trained 모델로 사용하였다. KR-ELECTRA 는 3 만 개의 vocab size 를 가진 한국어 형태소 분석기 Mecab-Ko 를 tokenizer 로 사용한다.
개인정보 특화 개체명 인식을 위해 tagging scheme 은 일반적으로 많이 사용하는 IOB2 포맷인 BIO 태그 방식을 사용하였고, tokenizer 가 분리한 token 각각을 개체명의 시작(Begin)과 개체명 내부(In), 개체명이 아닌 것(Out)으로 구분하였다.

모델 학습을 위한 epoch 은 20, batch size 는 32, learning rate 는 3e-5 로 설정하였고, optimizer 로는 AdamW 를 사용했으며, transformers 모듈의 get_cosine_schedule_with_warmup 이라는 scheduler 로 learning rate decay 를 실현하였다. 해당 모델로 일반 개체명 태그셋과 개인정보 특화 태그셋 실험을 진행하였다.

4.2 실험 결과

4.2.1 일반 개체명 태그셋과 개인정보 특화 태그셋 실험 결과

일반 개체명 태그셋과 개인정보 특화 실험 결과는 표 4.와 같다.

일반 개체명은 이름을 제외한 모든 항목에서 성능이 떨어지는 결과를 보였으며, 개인정보 별로 세부 개체명 분류가 가능했던 특화 개체명과 달리 5 개 대분류 개체명(PS, LC, OG, QT, TM) 밖에 사용하지 않아 범용적인 일반 개체명으로는 다양한 종류의 개인정보를 탐지할 수 없음을 확인하였다. 이에 개인정보적 관점에서 특화된 개체명 태그셋과 모델이 필요성을 확인할 수 있는 결과라 할 수 있다.

4.2.2 규칙 기반과 딥러닝 기반 실험 결과

개인정보 탐지 성능을 비교하기 위해 정규식을 활용한 규칙 기반 개인정보 탐지 성능과 개인정보 특화 개체명 태그셋을 활용한 KR-ELECTRA 모델의 개체명 인식률을 비교한 결과는 표 5.와 같다. 직장명은 정규식으로 탐지가 불가능하여 규칙 기반의 테스트에서는 제외하였다.

5. 결과 평가

본 연구는 실험 결과 데이터를 바탕으로 정량적 평가와 정성적 평가를 수행하였다. 정량적 평가는 규칙 기반과 딥러닝 기반의 precision, recall, f1-score 에 기반하여 분석하였으며 정성적 평가는 실험 결과에서의 오류 분석을 중심으로 진행하였다. 또한 정성적으로 오류를 분석한 결과를 토대로 개인정보 개체명의 보완 방향도 함께 살펴보았다.

5.1 정량적 평가

정규식을 통한 개인정보 탐지 성능은 이름과 주소, 직장명을 제외하고는 매우 우수한 성능을 보인다. 이는 개인정보에 해당하는 각각의 데이터가 정규식으로 정의 가능한 일반적인 형태로 이루어져 있기 때문이며, 실데이터에서는 '공 1 공-1234-56 칠 8'과 같이 정규식에 맞지 않은 형태의 변형된 개인정보에 대한 데이터들도 존재할 수 있기 때문에 규칙 기반의 개인정보 탐지 성능은 더욱 떨어질 수 있다.

표 4. 일반 개체명 태그셋과 개인정보 특화 태그셋 실험 결과 비교

	일반 개체명			개인정보 특화 개체명		
개인정보	Precision	Recall	F1	Precision	Recall	F1
이름	0.93	0.95	**0.94**	0.93	0.94	0.94
이메일	0.71	0.72	0.72	1.00	1.00	**1.00**
URL				0.96	1.00	0.98
주소	0.92	0.90	0.91	0.97	0.98	0.97
직장명	0.79	0.84	0.82	0.92	0.94	0.93
핸드폰번호	0.85	0.86	0.86	0.81	0.87	0.84
일반전화/FAX				0.97	0.99	0.98

표 5. 규칙 기반과 딥러닝 기반 실험 결과 비교

	rule-based			KR-ELECTRA		
개인정보	Precision	Recall	F1-score	Precision	Recall	F1-score
이름	0.02	1.00	0.03	0.93	0.94	0.94
이메일	1.00	1.00	**1.00**	1.00	1.00	**1.00**
URL	1.00	1.00	**1.00**	0.96	1.00	0.98
주소	0.04	0.99	0.07	0.97	0.98	0.97
직장명	-	-	-	0.92	0.94	0.93
핸드폰번호	1.00	1.00	**1.00**	0.81	0.87	0.84
일반전화/FAX	0.84	0.98	0.91	0.97	0.99	0.98

이름은 정규식으로 정의하기엔 무리가 있어 recall 은 높지만 f1-score 는 매우 저조한 것으로 나타난다. 주소는 딥러닝 모델에서 라벨링 된 시, 군, 구 등으로 분리된 형태와 마찬가지로 시/도, 군/구, 읍/면/동/가, 도로명 주소에 대한 정규식을 각각 적용하여 성능을 측정하였는데, “혹시”, “전화로”, “주셔도” 등과 같은 단어들까지 모두 주소로 예측하여

한편, KR-ELECTRA 모델의 개인정보 탐지 성능은 대체로 f1-score 와 recall, precision 의 편차가 크지 않은 결과를 보이며, 핸드폰번호를 제외한 모든 항목에서의 f1-score 는 0.93 이상의 수치를 보여준다. 정규식으로는 탐지가 어려웠던 이름과 주소의 f1-score 는 각각 0.94, 0.97 로 우수한 점수를 보이며, 정규식으로 탐지가 불가능했던 직장명 또한 f1-score 가 0.93 으로 잘 탐지해내는 것을 볼 수 있다.

f1-score 가 0.84 로 다른 항목들에 비해 상대적으로 낮은 핸드폰번호는 표 3-1.과 같이 학습용 데이터셋 볼륨이 낮기 때문인 것으로 보이며, 핸드폰번호에 대한 다량의 학습 데이터를 확보한다면 개선될 수 있을 것으로 보인다.

본 논문에서 진행한 실험에서는 데이터셋 생성 과정에서 개인정보에 해당하는 개체명들을 정규식 기반의 데이터들로 대치하였기 때문에 이메일, URL, 핸드폰번호 등의 분류가 잘 이뤄지는 것으로 보이지만 정규화되지 않은 형태의 데이터를 다수 포함한 실데이터에서는 성능이 떨어질 수 있다. 따라서 recall 이 높은 규칙 기반의 개인정보 탐지를 우선으로 수행한 뒤, 해당 결과에 대해 전반적으로 준수한 성능을 보이는 딥러닝 기반의 개인정보 탐지 모델을 통해 검증한다면 보다 우수한 결과를 보일 수 있을 것으로 기대한다.

5.2 정성적 평가

일반 개체명 실험 결과와 개인정보 특화 개체명 실험 결과를 각각 정성적으로 분석하였다. 일반 개체명 실험 결과의 오류를 분석하고 본 연구에서 제안하는 개인정보 특화 개체명이 범용적인 개체명 체계에서 발생하는 문제를 해결하는 데에 효율적인 체계가 될 수 있을지를 중심으로 하여 실험 결과를 분석하였다.

정성적 평가 결과 일반 개체명에서는 직장명과 연관된 ‘OG’와 관련된 오류가 가장 빈번하게 나타나는 것을 확인하였다. 자세한 오류 예시는 다음 표 6.과 같다. 예시의 주소와 직장명은 데이터 내의 내용을 그대로 옮겼으며, 이름과 번호류 유형은 비식별화하였다.

표 6. OG 관련 오류 예시

예시 1	주식회사 **에이치앤비디자인(OG)** OOO(PS)입니다.
예시 2	**필리아바이오(OG)**㈜에 연락 주세요. 친절히 안내해 드리도록 하겠습니다.
예시 3	**반갑습니다.** ㈜**핫마인드쓰리디(OG)** 상담원 OOO(PS)입니다.

위의 예시는 모두 ‘OG’의 범위 설정 오류로 분석할 수 있다. ‘㈜, 혹은 주식회사’까지 모두 포함되어 ‘OG’로 인식하여야 올바른 결과이다. 범용적으로 사용되는 일반 개체명으로는 직장명의 전체 스팬(span)을 포착하기에는 어려움이 있음을 알 수 있다. 정해진 패턴만을 ‘OG’에 적용할 수 없으며 ‘OG’ 표현 범위가 다양하게 나타나기 때문에 ‘OG’ 범위 설정의 오류가 다른 태그에 비해 빈번히 발생한다고 분석할 수 있다.

다른 유형의 오류로는 인명이 3 음절을 초과할 경우에 나타났다. 성과 이름이 4 음절 이상일 때에 이를 하나의 인명으로 인식하여야 하지만, 아래의 표 7.의 예시와 같이 ‘성’에 다른 태그로 잘못 할당하는 오류가 있었다. 이러한 오류는 한국어 인명의 전형적인 3 음절 패턴이 데이터의 대부분을 차지하고 데이터 내에서 희소하게 나타나는 3 음절 초과 인명은 데이터 구축에 있어 제대로 반영되지 않았기에 나타나는 결과로 해석할 수 있다. 특히 학습데이터 구축 시 다양한 음절의 이름에 대한 정확한 식별 결과를 얻기 위해서는 특이치(outlier)를 고려해야 할 것이다.

표 7. PS 관련 오류 예시

예시 1	북한산 밑에 있는 그 곳은 **O(OG)OOO(PS)** 촬영장소로 유명한 곳입니다.

일반 개체명 실험에서는 현재 사용되고 있는 도로명 주소를 하나의 주소 범위로 인식하지 못하고,

'LC+AF+QT'의 조합으로 쪼개서 인식하였다. 이 경우 'QT' 범위의 오류가 빈번하게 발견되었다. 이러한 오류는 위에서 분석한 'OG' 스팬 오류와 유사한 원인일 것으로 판단된다. 도로명 주소의 패턴이 다양하기 때문에 다양한 유형의 도로명 주소를 모두 정확하게 하나의 스팬으로 포착하지 못하였다.

표 8. OG, PS 관련 오류 유형 특화 개체명 적용 예시

예시 1	고객님 직접 방문하지 않고 **㈜엔유비즈(NUbiz Inc)(OG_WORKPLACE)**사이트를 통해 발급 받으실 수 있습니다.
예시 2	성함과 연락처 말씀해 주세요. 네, **OOOO(PS_NAME)**이고 OOO-OOOO-OOOO(QT_PHONE)입니다.

위의 표 8.은 앞서 살펴본 일반 개체명 실험 결과에서 나타난 오류 유형이 특화 개체명에서는 해결된 것을 보이는 예시이다. 'OG'와 'PS' 모두 올바른 범위로 인식된 것을 확인할 수 있다.

한편, 개인정보 특화 개체명 실험에서는 '님'을 인명의 범위로 인식하여 성을 제외한 'OO 님'을 인명으로 식별한 오류가 있었다. 이는 고객 상담을 주제로 한 말뭉치 도메인 특성상 높임을 표현하는 의존 명사 '님'이 빈번하게 나타나며 이를 이름과 띄어 쓰지 않고 붙어져 학습되어 이름으로 잘못 식별된 것으로 보여진다.

또한, 일반 개체명 태그셋으로 주석되어 있는 데이터를 개인정보 특화 개체명 태그셋으로 주석할 수 있는지에 대한 태그셋 적용 가능성을 확인하였다. 일반 개체명 데이터셋에서는 숫자로 표시된 대부분을 QT(Quantity)로 일반화하여 주석하였다. 개인정보 특화 개체명 태그셋을 적용하면 'QT'로 인식되었던 부분을 'LC_ADDRESS'로 주석함으로써 번지수까지 모두 하나의 주소로 인식할 수 있도록 할 수 있다.

표 9. 일반 개체명 태그셋의 개인정보 특화 적용 가능성 예시

	개체명 태그셋	**예시**
QT → LC	**일반 개체명 태그셋**	**경기도(LC) 하남(LC)시(LC) 덕풍북(AF)로(AF) 260(QT)** 주민(CV)입니다.
	개인정보 특화 개체명 태그셋	**경기도(LC_ADDRESS) 하남시(LC_ADDRESS) 덕풍북로(LC_ADDRESS) 260(LC_ADDRESS)** 주민입니다.

또한 일반 개체명 태그셋에서 전화번호 부분을 QT(Quantity)로 통일하여 제시하고 있으나 이를 개인정보 특화 개체명 태그셋으로 적용한다면 번호에 대해서 아래와 같이 '일반전화/FAX'와 '핸드폰 번호'로 세분화하여 주석할 수 있다.

표 10. 일반 개체명 태그셋의 개인정보 특화 적용 가능성 예시

	개체명 태그셋	**예시**
QT → QT_MOBILE	**일반 개체명 태그셋**	번호 123(QT)-4567(QT)요.
	개인정보 특화 개체명 태그셋	번호 012(QT_MOBILE)-3456(QT_MOBILE)-7890((QT_MOBILE)입니다.
QT → QT_PHONE	**일반 개체명 태그셋**	우리 사무소 123(QT)-456(QT)-7899(QT)로 문의 주세요.
	개인정보 특화 개체명 태그셋	우리 사무소 123(QT_PHONE)-456(QT_PHONE)-7899(QT_PHONE)로 문의 주세요.

위 예시들과 같이 기존의 일반 개체명에 개인정보 특화 개체명 태그셋을 적용하여 개체명을 더 세분화하여 주석할 수 있음을 확인하였다. 이처럼, 개인정보 특화 개체명 태그셋을 활용할 경우 개인정보 탐지가 용이하며 개인정보 비식별화 성능에 기여할 것으로 기대한다.

6. 결론

본 연구에서는 목적 지향적 대화를 원천 데이터로 하여 개인정보 탐지 및 비식별화를 위한 '개인정보 특화 개체명 데이터셋'을 구축하는 한편, 실험을 통해 일반 개체명 데이터셋의 개인 정보 분류 성능, 그리고 규칙 기반의 개인정보 탐지 성능보다 개인정보 특화 개체명 데이터셋을 사용할 경우 개인정보 탐지 분류 성능이 높아짐을 관찰하였다. 또한 정량 및 정성적 분석을 통해 개인정보 특화 개체명을 사용하여 개인정보를 주석할 경우, 정규식을 바탕으로 한 규칙 기반 개인정보 탐지의 한계점이었던 이름, 주소 등 일반 문자열로 구성된 개인정보에 대한 탐지가 가능해지는 것은 물론 일반 개체명에서 하나의 단위로 잡히지 않았던 비교적 긴 시퀀스의 정보에 대해서도 하나의 단위로 탐지될 수 있음을 확인하였다. 이를 통해 문맥 내 개인정보 탐지 시 개인정보를 주석할 수 있는 태그셋의 필요성을 환기하는 한편 실제 산업 현장에서 개인정보 특화 개체명 태그셋을 활용할 수 있는 가능성 역시 확인하였다. 다만 본 연구는 기초 단계인 바, 개인정보 특화 태그셋을 7 개만 사용하였고 원천 데이터 역시 일상 대화(chit-chat)가 아닌 목적 지향 대화인 점, 그리고 유출 위험 정도(intensity)에 대한 세밀한 주석이 이루어지지 않은 한계가 있었다. 이에 따라 향후 연구에서는 일상 대화에 대해서도 연구를 진행하는 한편, 추가적인 주석 라벨 세밀도(granularity)나 라벨별 유출 위험 정도(intensity)를 연구할 예정이다.

감사의 글

본 연구는 2022 년도 정부(개인정보보호위원회)의 재원으로 한국인터넷진흥원의 지원을 받아 수행된 연구임(No. 1781000006, 대화형 텍스트 데이터에서 AI 기반 개인정보 탐지 및 비식별화 기술 개발)

참고문헌

[1] Campanile, Lelio, et al. "Sensitive Information Detection Adopting Named Entity Recognition: A Proposed Methodology." International Conference on Computational Science and Its Applications. Springer, Cham, 2022.

[2] 박지성, “Personal sensitive information identification through deep learning based named entity recognition in structured documents”, 국내석사학위논문 한양대학교 대학원, 2020.

[3] 민연아, “데이터 사용 범위를 고려한 가명정보 처리방법 연구”, 한국컴퓨터정보학회논문지 ,26(5),17-22, 2021.

[4] 서동국, 김건우, 김재영, 이동호. "문장 의도 분류와 개체명 인식을 활용한 개인정보 검출 및 비식별화 시스템." 한국정보처리학회 학술대회논문집 27.2 ,1018-1021, 2020.

[5] 최재훈, 조상현, 김민호, 권혁철, “개인정보 비식별화를 위한 개체명 유형 재정의와 학습데이터 생성 방법”, 한국정보통신학회 종합학술대회 논문집,26(1),206-208, 2022.

[6] 서용호. "기계학습을 활용한 개인정보 식별율 향상에 관한 연구." 국내석사학위논문 숭실대학교 대학원, 2019.

[7] Bannour, N., Wajsbürt, P., Rance, B., Tannier, X., & Névéol, A. (2022). Privacy-preserving mimic models for clinical named entity recognition in French. Journal of Biomedical Informatics, 130, 104073, 2022.

[8] G. B. Herwanto, G. Quirchmayr and A. M. Tjoa, "A Named Entity Recognition Based Approach for Privacy Requirements Engineering," 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW), 2021, pp. 406-411.