

개인정보 특화 개체명 주석 대화 데이터셋 기반

생성시 활용 개체명 탐지

강예지^o, 비립, 장연지, 박서윤, 김한샘[†]

연세대학교 언어정보학협동과정[†]

{ yjkang5009^o, feili0820, yeonji3547, seoyoon.park, khss[†] }@yonsei.ac.kr

Named Entity Detection Using Generative AI for Personal Information- Specific Named Entity Annotation Conversation Dataset

Yejee Kang^o, Li Fei, Yeonji Jang, Seoyoon Park, Hansaem Kim[†]

Yonsei University, Interdisciplinary Graduate Program of Linguistics and Informatics

요약

본 연구에서는 민감한 개인정보의 유출과 남용 위험이 높아지고 있는 상황에서 정확한 개인정보 탐지 및 비식별화의 효율을 높이기 위해 개인정보 항목에 특화된 개체명 체계를 개발하였다. 개인정보 태그셋이 주석된 대화 데이터 4,981세트를 구축하고, 생성 AI 모델을 활용하여 개인정보 개체명 탐지 실험을 수행하였다. 실험을 위해 최적의 프롬프트를 설계하여 퓨샷러닝(few-shot learning)을 통해 탐지 결과를 평가하였다. 구축한 데이터셋과 영어 기반의 개인정보 주석 데이터셋을 비교 분석한 결과 고유식별번호 항목에 대해 본 연구에서 구축한 데이터셋에서 더 높은 탐지 성능이 나타났으며, 이를 통해 데이터셋의 필요성과 우수성을 입증하였다.

주제어: 개인정보 특화 개체명, 개인정보 탐지, 생성시, 프롬프트 설계

1. 서론

대량의 데이터가 수집, 저장, 그리고 공유되는 현 시대에서 민감한 개인정보의 유출이나 남용의 위험이 증가하고 있다. 개인정보 유출은 개인뿐만 아니라 조직, 기업, 공공기관에도 심각한 영향을 미칠 수 있으며 이로 인해 금전적 손실은 물론 법적 문제, 규제 준수 요구사항 등과 관련해서도 큰 영향을 미치고 있다. 따라서 정확한 개인정보 탐지가 핵심적인 과제로 부각되고 있으며, 이를 통해 민감한 정보의 식별 및 보호가 강화되어야 한다. 민감 정보를 안전하게 보호하기 위해 이를 비식별화하는 마스킹 처리 과정을 거쳐야 하는데, 마스킹 처리를 위해서는 개인정보를 정확하게 탐지하는 과정이 반드시 선행되어야 한다.

데이터 내 개인정보를 정확하게 탐지하고 이를 마스킹하기 위해 개인정보의 유형을 정의하고 개인정보 항목에 특화된 개체명 체계가 필요하다. 이는 일반 개체명이 주석된 데이터만을 사용할 경우 개인정보의 완전한 비식별화를 보장하기 어렵기 때문이다.

본 연구는 개인정보를 정확히 탐지하고 비식별화하기 위해 개인정보 특화 개체명 체계를 고안하고, 개인정보 특화 개체명이 주석된 대화 데이터 총 4,981세트를 구축하였다. 구축한 개인정보 태그셋 데이터셋을 대상으로 생성 AI 모델을 활용하여 탐지 실험을 진행하였다. 실험을 위해 모델이 개인정보 특화 개체명 탐지 태스크를 수행하도록 프롬프트를 설계하였다. 모델은 대규모 모델과

경량화 모델을 이용하여 그 결과를 비교 분석함으로써 탐지 성능을 평가하였으며, 본 연구에서 구축한 데이터셋 외 영어 기반의 개인정보 주석 데이터셋을 대상으로 실험을 진행하여 구축한 데이터셋의 필요성 및 우수성을 검증하였다.

2. 관련 연구

2.1. 개인정보 관련 개체명 인식 및 비식별화 연구

개인정보와 관련한 개체명을 재정의하여 사용한 연구로는 [1]~[4]가 있다. [1]은 개인정보 관련 개체명을 새롭게 정의하고 이를 기반으로 한국어 위키피디아 본문을 활용하여 비식별화된 데이터를 생성하는 형식을 제안하였다. 개체명의 경우 12개의 유형으로 구분하여 ‘사람 이름, 별명, 캐릭터, 가수/그룹명’과 같이 개인을 식별하는 정보를 포함할 수 있도록 하였으며, ‘혈액형, 병명, 전공, 학력’ 등과 같은 정보의 경우 규칙에 따라 수작업으로 부착하였다. 실험에서 모델이 개체명을 인식하지 못할 경우 해당 개체명 태그는 삭제한 후 반자동 시스템을 통해 수정할 수 있도록 하였다. 한편 기존의 개인정보 탐지 방법으로 사용된 패턴 및 형태소 분석 기법의 한계를 지적한 [2]는 기계학습 알고리즘을 이용하여 개인정보 인식율을 향상시키는 방법을 제시하였다. 이를 위해 개인정보로 인식될 수 있는 정보를 개인정보 비식별화 대상으로 삼고 개인정보 태그셋을 구축하였다. 이후 데이터를 이용하여 다양한 패턴

과 유형을 학습한 후 기계학습 알고리즘에 변화를 주어 개인정보를 추출하는 실험을 반복하였다. 그 결과 패턴 기반의 개인정보 인식 방식보다 기계학습 기법을 사용한 반복적으로 재학습했을 경우 인식율이 더 향상된 것을 확인하였다.

[3]과 [4]에서는 특수 도메인을 대상으로 개인정보 탐지를 시도하였다. [3]은 법률 도메인에서 문서 내 개체를 식별하고 민감한 정보를 추출해내기 위해 generic NER을 사용하였다. 식별된 개체는 적절한 컨텍스트 윈도우에 포함시켜 동일한 도메인의 텍스트에서 비슷한 단어와 얼마나 유사한지를 평가하였다. [4]에서는 특수 도메인인 의료 분야 내 개인정보 개체명을 탐지하기 위해 여러 프랑스어 의료 분야 데이터셋을 사용하였으며, 각 데이터셋에서 나타나는 개체명 태그를 통합하였다. 개인정보 관련 개체명을 탐지하기 위해 teacher-student 모델을 활용하였고, 실험 결과 개인정보 개체명 데이터셋 구축의 필요성과 자동 주석의 가능성을 확인하였다.

개인정보 개체명 자동 주석과 관련한 연구는 [5]에서도 나타난다. [5]는 위키피디아 전기(Wikipedia biography) 페이지를 대상으로 개인정보 추출을 위해 WikiPII라는 자동 주석 데이터셋을 구축하였다. 데이터는 BERT 기반 개체명 인식 모델을 이용하여 학습하였으며, 대량의 학습 데이터셋을 사용하면 라벨의 노이즈가 높더라도 수동 추출 비용을 크게 줄일 수 있음을 보여주었다. [6]은 문장의 의도 정보 분류 결과를 개체명 학습에 활용함으로써 개체명 인식 결과만을 사용했을 때 개인정보 외 정보에도 비식별화 처리가 되는 한계점을 극복하고자 하였다. 그 결과 문장 의도 분류를 학습 시 추가 정보로 사용한 경우 성능이 더 향상된 것을 확인하였다.

한편, 본 연구의 기초연구였던 [7]은 목적 지향적 대화를 원천 데이터로 하여 개인정보 특화 개체명 데이터셋을 구축하였으며 일반 개체명 데이터셋과의 비교 실험을 통해 개인 정보 탐지 성능을 확인하였다. 또한 개인정보 특화 개체명 데이터셋을 사용할 경우 규칙기반의 개인정보 탐지 성능보다 높아짐을 확인함으로써 개인정보 태그의 필요성을 환기하였다. 하지만 원천 데이터가 목적 지향적 대화이며, 개인정보 특화 태그셋이 7개로 그 수가 제한적이었다는 점에서 한계가 있었다.

2.2. 프롬프트 엔지니어링

프롬프트 엔지니어링은 다운스트림 작업에서 가장 효과적인 성능을 얻기 위한 프롬프트 함수를 생성하는 과정으로, 모델이 수행할 것으로 예상하는 작업에 대해 가장 적합한 템플릿을 찾는 과정을 말한다[8]. 프롬프트에는 주어진 텍스트 문자열의 빈칸을 채우는 방식으로 동작하는 클로즈(Cloze) 프롬프트 유형과 문자열의 접두사를 이어서 동작하는 프리픽스(Prefix) 프롬프트 유형이 있다.

프롬프트를 생성하는 가장 자연스러운 방법 중 하나는 인간이 직접 직관적인 템플릿을 수동으로 만드는 것이다.

[9]의 LAMA 데이터의 경우 수동으로 생성한 클로즈 템플릿을 제공하고, [10]은 질문 응답, 번역 및 상식적 추론을 위한 다양한 태스크를 수행하기 위해 프리픽스 프롬프트를 수동으로 생성하였다. 이처럼 각 태스크에 맞게 정의된 템플릿을 사용하여 학습 환경에 활용한다.

본 연구에서는 기존의 개인정보 관련 개체명 인식 및 비식별화 연구의 한계를 극복하고자 개인정보 특화 개체명을 새롭게 정의하고 이를 기반하여 일상 대화를 주제로 하는 개인정보 개체명 주석 데이터셋을 구축하고자 하였다. 또한 개인정보 특화 개체명 탐지라는 다운스트림 태스크를 수행하기 위해 최적의 프롬프트를 수동으로 설계하였으며, few-shot 학습 환경에 맞게 구성하였다.

3. 개인정보 특화 개체명 주석 대화 데이터셋

3.1. 대화 데이터셋 구축

대화형 텍스트 데이터에서 AI를 기반으로 개인정보를 탐지하기 위해서는 개인정보가 포함된 일상 대화 데이터셋을 구축하는 작업이 필요하다. 본 연구에서는 AIHub의 ‘한국어 SNS’ 데이터에서 사용한 주제 분류를 참고하여 8개의 대화 주제(개인 및 관계, 주거와 생활, 쇼핑과 거래, 공공 서비스, 여가와 오락, 일과 직업, 미용과 건강, 배움과 진로)를 중심으로 개인정보를 포함할 수 있도록 다양한 주제의 대화 문장을 구성하였다.

구체적으로 하나의 대화 세트는 두 명의 화자의 대화로 한정하며, 최소 3턴 그리고 평균 4턴 이상으로 구성한다. 구축 시 다수를 대상으로 하는 이야기에 대한 대화는 작성하지 않고, 대화 중간에 제3자가 끼어들거나 시간 공백이 발생하는 대화는 작성하지 않도록 한다. 또한 이모티콘의 지나친 남용은 허용하지 않으며, 비속어, 은어, 사투리, 맞춤법 등은 TV 방송에서 허용되는 수준까지 사용할 수 있도록 한다. 마지막으로 개인정보 항목에 해당하는 성별, 몸무게, 키, 나이 등과 관련한 내용을 제외하고는 차별적이거나 혐오적인 표현은 사용하지 않도록 한다.

구축된 대화 세트는 각 세트마다 대화 주제를 분류하여 표기되되 주제가 명확하게 구분되지 않을 경우 ‘개인 및 관계’로 분류하였다. 위와 같은 원칙을 바탕으로 대화 데이터 총 4,981세트, 10,900개의 문장을 구축하였다¹.

3.2. 개인정보 특화 개체명 주석 가이드라인

개인정보 항목의 대분류는 8가지로, 개인 일반, 위치 정보, 고유식별번호, 일반식별정보, 직장 정보, 학교 정보, 온라인 정보, 군사 정보로 이루어져 있다. 8가지의 대분류 항목은 아래 표와 같이 항목별로 총 33가지의 세분류 개체명으로 분류된다.

¹ 1차 데이터 구축은 (주)스피링크에 의뢰하여 수행하였다.

표 1 개인정보 항목 및 개체명 태그셋

	개인정보 항목	개인정보 특화 개체명	
1	개인 일반	이름	PS_NAME
2		별명/애칭	PS_NICKNAME
3		생년월일	DT_BIRTH
4		나이	QT_AGE
5		성별	CV_SEX
6		키	QT_LENGTH
7		몸무게	QT_WEIGHT
8		혈액형	TM_BLOOD_TYPE
9		종교	OGG_RELIGION
10		국적	LCP_COUNTRY
11		동아리/동호회	OGG_CLUB
12	위치 정보	주소	LC_ADDRESS
13		장소명	LC_PLACE
14	고유식 별번호	주민등록번호	QT_RESIDENT_NUMBE R
15		외국인번호	QT_ALIEN_NUMBER
16		여권번호	QT_PASSPORT_NUMBE R
17		운전면허번호	QT_DRIVER_NUMBER
18	식별정 보(일 반)	핸드폰번호	QT_MOBILE
19		일반전화/FAX 번 호	QT_PHONE
20		카드번호	QT_CARD_NUMBER
21		계좌번호	QT_ACCOUNT_NUMBER
22		이메일주소	TMI_EMAIL
23	직장 정보	차량번호	QT_PLATE_NUMBER
24		직장명	OG_WORKPLACE
25		부서명	OG_DEPARTMENT
26	학교 정보	직책/직급	CV_POSITION
27		학교명	OGG_EDUCATION
28		학년	QT_GRADE
29	온라인 정보	전공	FD_MAJOR
30		ID	PS_ID
31		URL	TMI_SITE
32	군사 정보	IP 정보	QT_IP
33		근무부대	CV_MILITARY_CAMP

개인정보 특화 개체명 태깅 대원칙은 다음과 같다².

<p>1. 최장 개체명을 기준으로 태깅한다. 예) <한국해양대학교:OGG_EDUCATION> 0 <한국:LCP_COUNTRY> <해양 대학:OGG_EDUCATION> X</p>

<p>예외) 각각의 명사 정보를 참조할 필요가 있는 경우 ‘최소 단위 주석’ 을 기준으로 하여 각 단어를 따로 태깅한다. 예) <부산광역시 해운대구 좌동:LC_ADDRESS> X <부산광역시:LC_ADDRESS><해운대구:LC_ADDRESS> <좌동:LC_ADDRESS> 0</p> <p>2. 줄임말이나 단어를 변경하여 개체명을 표현할 경우 원 단어를 기준으로 태깅한다.</p> <p>3. 하나의 개체명이 여러 개의 개인정보 항목에 해당 될 경우 맥락을 기반으로 태깅한다.</p> <p>4. 음차 표기는 제외한다. 예외) 생년월일, 나이, 키, 몸무게의 숫자 표기는 음차 표기를 허용하며, 함께 나오는 단위 역시 음차 표기를 허용한다. 또한 외국인과 관련한 이름, 주소/장소명, 학교/직장명은 음차 표기를 허용하며 이때 외래어 표기법에 맞춰 표기한다.</p> <p>5. 정확한 수치가 아닌 어림짐작 등의 표현, 미래에 대한 개인정보는 태깅 대상으로 삼지 않는다.</p>
--

이때 개인정보 특화 개체명의 경우 실제 대화에서 자연스럽게 가장 많이 쓰일 법한 개체명의 구축 비율을 높이고, ‘고유 식별 번호’, ‘IP정보’ 등과 같이 특수 상황에서만 언급되는 개체명의 경우 최소 200개는 확보할 수 있도록 구축하였다. 최종적으로 태깅 대원칙에 따라 총 4,981개의 대화 세트를 대상으로 개인정보 특화 개체명은 23,808개가 주석되었다. 아래 표는 개인정보 특화 개체명을 주석한 대화 세트 예시이다.

표 2 개인정보 특화 개체명 주석 대화 세트 예시

대화 번호	도메인 구분	발 화 자	발화문
1	개인 및 관계	P01	언니, 다 끝났으면 얼른 주차장으로 와.
2		P01	나 <01 어 2902:QT_PLATE_NUMBER> 흰색 차야.
3		P02	어, 잠시만... 나 지갑을 잃어버린 거 같아.
4		P01	마지막에 어디서 봤는데?
5		P02	아까 <스타벅스 논현점:LC_PLACE>에서 커피 계산했으니까 그때는 확실히 있었고...
6		P02	화장실에 두고 왔나?
7		P01	일단 화장실 먼저 가보고 가면서 분실센터에도 전화해 봐.

²지면상의 한계로 자세한 주석 가이드라인은 생략한다.

8	P01	전화번호 <052-736-9163:QT_PHONE>이야.
9	P02	응, 그래야겠다.

4. 생성AI를 이용한 개체명 탐지 실험

생성AI를 이용한 개인정보 개체명 탐지 실험은 3장에서 구축한 데이터셋의 필요성과 우수성 두 가지 측면을 검증하기 위해 진행되었다. 먼저 생성형 AI 모델의 개인정보 개체명 탐지 실험을 통해 모델의 개인정보 개체명에 대한 이해도를 판단할 수 있다. 모델이 개인정보 개체명에 대한 탐지를 잘 수행하지 못한다면 구축한 데이터가 학습데이터로서 의의가 있을 것이다. 두 번째로 허깅페이스에 공개된 개인정보 개체명이 주석된 영어 데이터³를 이용하여 두 데이터의 개체명 탐지 결과를 비교하여 구축한 데이터셋의 우수성을 보장하였다.

개인정보 특화 개체명 실험에는 현재 공개되어 있는 생성AI 모델 중 이용자 수가 가장 많은 ChatGPT와 경량화 모델 Llama2 모델을 사용하여 개체명 탐지 결과를 비교하였다.

개인정보 탐지 실험은 식별번호 항목을 중심으로 숫자로만 구성되어 개체명 간 구분이 어려운 고유식별번호 중 ‘주민등록번호, 여권번호’와 일반 식별정보 중 ‘핸드폰번호, 일반전화/FAX 번호, 카드번호, 계좌번호’ 총 6개 개체명에 대해 진행하였다. 아래 표는 실험에 사용된 식별번호 개체명과 그 예시이다.

표 3 고유식별번호 개체명 및 발화문 예시

개체명 항목	개체명 주석 예시
주민등록번호	알아요, 엄마는 주민등록번호 <760726-2329082:QT_RESIDENT_NUMBER>이거예요.
여권번호	저 여권번호가 <M473U04003:QT_PASSPORT_NUMBER>인데 이거 어디다 적어야 하는 거예요?
핸드폰번호	지금 전화주신 <010-3326-3335:QT_MOBILE> 번호의 손진향 고객님의 본인 맞으실까요?
일반전화/FAX 번호	전화로 가능하신데, 그건 저희 쪽이 아니라 <054-6458-4873:QT_PHONE>으로 전화주셔야 해요.
카드번호	<부산비씨 3793 505479 13564:QT_CARD_NUMBER>입니다.
계좌번호	장학금 받으실 계좌 번호가 <500898-40-773917 우체국:QT_ACCOUNT_NUMBER> 그대로이실까요?

4.1. 프롬프트 설계

먼저 7개의 개인정보 개체명을 탐지하기 위한 프롬프트를 설계하였다. 실험에 사용된 프롬프트는 다음과 같다.

주어진 문장에 대해서 개인정보 개체명 분석을 수행하려고 해.

(1) 개인정보 항목: 주민등록번호, 외국인번호, 여권번호, 핸드폰번호, 일반전화/FAX번호, 카드번호, 계좌번호

(2) 개인정보 항목 개체명 태그셋:

주민등록번호 <QT_RESIDENT_NUMBER>

여권번호 <QT_PASSPORT_NUMBER>

핸드폰번호 <QT_MOBILE>

일반전화/FAX 번호 <QT_PHONE>

카드번호 <QT_CARD_NUMBER>

계좌번호 <QT_ACCOUNT_NUMBER>

개인정보 항목에 대해 <>로 표시해줘.

=====

분석 예시:

[주민등록번호]

- 알아요, 엄마는 주민등록번호 <760726-2329082:QT_RESIDENT_NUMBER>이거예요.

- 내거는 <100605-4703713:QT_RESIDENT_NUMBER>, 아빠 건 <670605-1703713:QT_RESIDENT_NUMBER>.

[여권번호]

- 저 여권번호가 <M473U04003:QT_PASSPORT_NUMBER>인데 이거 어디다 적어야 하는 거예요?

- 영문 이름도 넣었고 여권번호도 <M35850359:QT_PASSPORT_NUMBER>이고.

- 이거 <M44319051:QT_PASSPORT_NUMBER>을 세 번째 칸에 넣는거지?

- 오른쪽 위에 <M63490189:QT_PASSPORT_NUMBER> 맞아?

- 우리 여권번호 <M94163099:QT_PASSPORT_NUMBER>누구야?

...

[계좌번호]

- 장학금 받으실 계좌 번호가 <500898-40-773917 우체국:QT_ACCOUNT_NUMBER> 그대로이실까요?

- 아냐 내가 할테니까 <시티 770-524252-430-68:QT_ACCOUNT_NUMBER>로 반만 보내줘.

- 저는 어제 술 드신거 제 카드로 결제했으니까 <3995 11 7718117 카카오:QT_ACCOUNT_NUMBER>에 반 보내주시고요.

- 제가 이거 바잉 걸어둘테니까 <기업은행 223-583353-76-703:QT_ACCOUNT_NUMBER>로 입금 천천히 하시구요.

=====

문장:

그럼 보내고 나서 01091432312로 문자 하시든지 아니면 해피 홈부동산 1522-2383로 전화하세요.

너 여기 등록할 때 M69099309 저걸로 했잖아.

6개의 개인정보 항목 개체명과 태그셋을 넣고 그 아래 입력값과 출력값을 각각의 항목별 예시로 추가하였다. 이때 각 개인정보 항목을 구분하지 않고 통으로 예시를 넣는 경우보다 ‘[]’로 구분하여 예시를 제시하는 것이 결과문을 항목별로 더 잘 구분해냈다.

4.2. 퓨샷 러닝

프롬프트에 예시로 들어가는 최적의 샷을 찾기 위해 샷의 개수를 1, 3, 7개로 서로 다르게 설정하여 같은 10개

³ ai4privacy/pii-masking-65k

의 입력값에 대해 그 출력값의 정확도를 계산하였다.

이때 본 연구에서 구축한 데이터의 우수성을 검증하기 위해 허깅페이스에 공개된 영어 기반의 개인정보 개체명 주석 데이터를 이용하여 탐지 실험을 진행하고 그 결과를 비교하였다.

5. 평가

5.1. 샷 개수에 따른 모델의 탐지 결과

4.2절에서 언급한 바와 같이 샷의 개수를 1, 3, 7개로 다르게 설정하여 10개의 입력문(개체명 10개)에 대한 출력값을 비교한 결과는 아래와 같다.

표 4 샷의 개수에 따른 개인정보 개체명 탐지 결과

shot	ChatGPT	Llama2
1	10/10	9/10
3	10/10	2/10
7	9/10	1/10

샷의 개수가 1, 3개일 때, ChatGPT의 경우 아래 그림과 같이 고유식별번호를 잘 구분하여 탐지하고 있음을 확인할 수 있다. 반면 샷의 개수를 7개로 늘렸을 때는 “주민등록번호” 항목에 대한 탐지가 제대로 이루어지지 못하였다.

구는 추후 진행되어야 할 것이다.

```
Llama 2
Sure, I can analyze the given sentence and highlight the personal information items with <> brackets.
Here's the sentence:
"나는 010-1234-5678로 전화하고 싶어요. 제 이름은 김나나입니다. 주민등록번호는 760726-2329082입니다. Currently, I am living in 서울시 gangnam-gu, 삼성로 321-45 edificio, 402호예요. My passport number is M473U04003, and my phone number is 010-3326-3335. I have a 부산비씨 3793 505479 13564 card, and my account number is 500898-40-773917 우체국. I would like to receive the phone call at 14:30 on November 23rd."
Here are the personal information items:
• Phone number: 010-1234-5678
• Name: 김나나
• Resident registration number: 760726-2329082
• Address: 서울시 gangnam-gu, 삼성로 321-45 edificio, 402호
• Passport number: M473U04003
• Phone number: 010-3326-3335
• Card number: 부산비씨 3793 505479 13564
• Account number: 500898-40-773917 우체국
I have highlighted the personal information items with <> brackets for easy identification.
```

그림 2 Llama2 오탐지 결과(3-shot)

5.2. 데이터에 따른 모델의 탐지 결과

본 연구에서 구축한 데이터와 기주석된 개인정보 개체명 주석 데이터셋의 탐지 결과를 비교하고자 허깅페이스에 공개된 pii-masking-65k 데이터셋을 사용하였다. Pii-masking-65k 데이터셋은 AI assistant 및 LLM 모델의 대화 맥락의 텍스트에서 개인식별정보(PII)를 제거하기 위해 구축된 영어 기반 데이터로, 탐지 비교 실험을 위해 한국어로 번역하여 진행하였다. 실험에는 ‘카드번호’, ‘전화번호’를 비롯해 온라인 정보 중 하나인 ‘IP정보’를 추가하여 총 세 가지 항목을 사용하였다.

프롬프트의 경우 4.1절에서 설계한 것과 동일하며 샷의 수는 1, 3개로 설정하였다. 또한 Llama2 모델의 경우 유의미한 결과가 나오지 않아 ChatGPT 만 사용하여 분석하였다.

표 5 데이터에 따른 개인정보 개체명 탐지 결과

shot	Our dataset	Pii-masking-65k
1	10/10	9/10
3	10/10	8/10

그 결과 표 5와 같이 본 연구에서 구축한 데이터셋의 경우 숫자로만 구성된 카드번호, 전화번호, IP 정보 개체명에 대한 탐지에 성공하였다. 반면, pii-masking-65k의 번역 데이터의 경우 아래 예와 같이 전화번호를 미탐지하거나 개인정보 관련 숫자가 아닌 일반 숫자 역시 개인정보로 탐지하는 오류가 있었다.

표 6 pii-masking-65k 데이터셋 개체명 탐지 오류 예시

- 업무 중단 시 살바토레 와이즈낫이 사용할 수 있도록 Rhea26@hotmail.com 및 956-260-5936 x61190을 연락처 정보로 포함하는 커뮤니케이션 계획을 수립하세요

이 문장에는 개인정보 항목이나 IP 정보가 포함되어 있지 않습니다.

주어진 문장에서 개인정보 개체명을 분석한 결과는 다음과 같습니다:

- 그림 보내고 나서 <01091432312:QT_MOBILE>로 문자 하시든지 아니면 해피홈부동산 <1522-2383:QT_PHONE>로 전화하세요.
- 너 여기 등록할 때 M69099309:QT_PASSPORT_NUMBER 저걸로 했잖아.
- 그럼 요금은 <3762 566089 75895:QT_CARD_NUMBER>로 납부할게요.
- 부산비씨 <3793 505479 13564:QT_CARD_NUMBER>입니다.
- <010.7762.4678:QT_MOBILE> 인데 나 지금 바빠죽겠으니까 얼른 처리해주세요.
- 네, 신규가입부 번호는 <02-8961-3909:QT_PHONE>입니다.
- 속소 배정이 확정되면 <107-789-155325:QT_ACCOUNT_NUMBER> 신함으로 입금해 주시면 되고, 신청해 주신 2인실은 한 학기에 총 220만 원이에요.
- 내가 <911121-1417643:QT_RESIDENT_NUMBER>이고, 마누라가 <910124-2651228:QT_RESIDENT_NUMBER>...

그림 1 ChatGPT 탐지 결과(1-shot, 3-shot)

한편, Llama2의 경우 샷을 1개 넣었을 때 고유식별번호를 가장 잘 탐지하였으며, 샷의 개수를 3개와 7개로 늘렸을 때 오히려 탐지 성능이 떨어짐을 확인하였다. 샷을 1개로 설정하였을 경우 ‘여권번호’를 ‘주민등록번호’로 잘못 인식하는 문제가 있었으며, 샷을 늘렸을 때는 주어진 입력값을 바탕으로 출력값을 도출해 내는 것이 아닌 그림 2와 같이 제3의 문장을 생성하여 문장 내 개인정보를 탐지하는 결과를 보였다. 또한 Llama2 모델의 경우 출력문에 한국어와 영어가 혼재되어 나타났으며 개인정보 항목에 대한 이해가 떨어지는 것을 확인할 수 있었다. 이는 경량화 모델임에도 불구하고 ChatGPT-3.5보다 더 높은 성능을 보인다는 Meta AI의 리포트를 뒷받침하지 못하였다. ChatGPT의 경우 Llama2보다 학습된 데이터 규모가 크다는 점에 기인하여 모델의 성능이 학습 데이터 양에 영향을 미치는 것으로 보여진다. 한편 샷의 개수가 많아질수록 탐지 성능이 떨어지는 것에 대한 연

