대화형 텍스트 데이터 내 개인정보 식별에 대한 연구

차도현 [1, †], 권보근 [1, †], 윤희창 [1, †], 이구협 [1, †], 주종화 [1,*]
[1] 동국대학교 컴퓨터정보통신공학부 컴퓨터공학전공
cha2016112155@gmail.com[1, †], rnjsqhrms123@gmail.com[1, †], hcfree1500@gmail.com[1, †], glen2003@dgu.ac.kr[1, †], jwjjoo@dgu.ac.kr[1, *]
[†]공동 1 저자
*교신저자
[1] 동국대학교 컴퓨터공학과, 04620 서울, 대한민국

A Study on Identifying Personal Information on Conversational Text Data

Do Hyun Cha[1, †], Bo Keun Kown[1, †], Hee Chang Youn[1, †], Gu Hyup Lee[1, †], Jong Wha J. Joo[1,*]
[1]Department of Computer Science and Engineering, Dongguk University-Seoul, 04620 Seoul, South Korea
[†] Do Hyun Cha, Bo Keun Kwon, Hee Chang Youn, Gu Hyup Lee contributed equally to this work.
* Correspondence: jwjjoo@dgu.ac.kr

요 약

데이터 3 법을 필두로, 기업은 개인정보가 포함된 데이터를 활용하기 위해 비식별 처리가 필요하게 되었다. 기존 방식은, 비정형 텍스트 데이터에서 정규표현식을 통한 개인정보 식별은 데이터의 다양성에 의해 한계가 명확하며, 기존의 Named Entity Recognition(NER) 태스크로 해결하기에는 언어의 중의적 표현과 2 인 대화에서 나타나는 개인정보가 누구의 것인지 판단하지 못한다는 한계가 존재한다. 따라서 우리는 기존의 한계점을 극복하고 개선하기 위해 BERT 언어 모델에 화자 정보를 학습시키고, 하나의 어절에 2 개의 tag 를 labeling 하는 방법을 제안하여 정확한 개인정보 식별을 시도하였다.

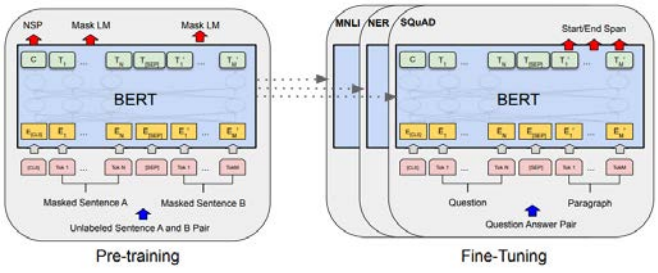
1. 서론

2020 년 08 월 개인정보 보호법, 정보통신망법, 신용정보법으로 구성된 일명 데이터 3 법 개정안이 시행되었다. 가명정보 개념을 도입하여 주체의 동의 없이 데이터 활용이 가능해졌고, 이 개정안으로 인해 데이터 활용 창구는 늘어났으며, 개인정보를 식별하여 비식별화하는 기술의 중요성이 대두되었다.

본 연구에서는 이러한 개인정보를 비식별하기 위해 대화 데이터에서 개인정보를 식별, 해당 개인정보가 누구의 개인정보인지 파악할 수 있는 방법을 제안한다.

기존의 딥러닝을 활용한 자연어 처리는 BERT[1], ALBERT[2], RoBERTa[3]와 같이 트랜스포머 기반의 사전 학습 언어 모델을 활용해 여러 태스크에서 괭장한 성능 향상을 보이고 있다. BERT 는 bidirectional 언어 모델로, 문맥을 양방향으로 보고 문장 내에서 마스킹된 임의의 단어를 예측하는 Masked LM(MLM) 태스크와 문장 A 와 B 가 입력되었을 때 문장 B 가 문장 A 의 다음 문장으로 적절한지 판단하는 Next Sentence Prediction(NSP) 태스크에 대해 모델을 학습한다. [그림 1]은 BERT 모델의 구조를 표현한 것이다.

본 연구에서는 사전 학습된 BERT 모델에서 해당 개인정보가 누구의 개인정보인지 파악하기 위한 기법들을 제안한다. SNS 와 같은 구어체 데이터로 사전 학습된 한국어 언어 모델 KcBERT 에 이러한 기법을 적용하였고 누구의 개인 정보인지 식별할 수 있도록 개선하였다.



<그림 1> BERT Language Model 의 구조

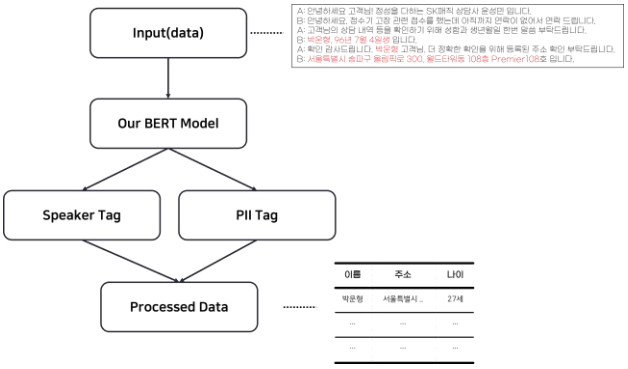
2. 관련 연구

최근 많은 개인정보 식별 연구는 정규 표현식과 Named Entity Recognition(NER)로 해결하고 있다. 특히 한국어로 학습된 BERT 모델인 KoBERT[4]는 NER 태스크에서 매우 높은 정확도를 갖는다. 그러나 이런

시도는 중의적 표현을 제대로 식별할 수 없고, 식별된 개인정보가 발화자 중 누구의 개인정보인지 확인하기 어렵다.

NER 태스크는 하나의 개체명에 인명, 단체, 장소와 같은 속성값(이하 태그) 중 하나를 라벨링하는 작업이다. 여러 연구에서 이러한 작업을 위해 언어 모델 출력층에 하나의 선형 레이어와 시그모이드 활성화 함수를 추가하고 소프트맥스 함수를 통해 어절을 예측한다. 이러한 모델 구조는 하나의 어절에 여러 개의 태그를 달 수 없는 한계가 존재한다.

3. 대화형 데이터 기반 BERT 모델

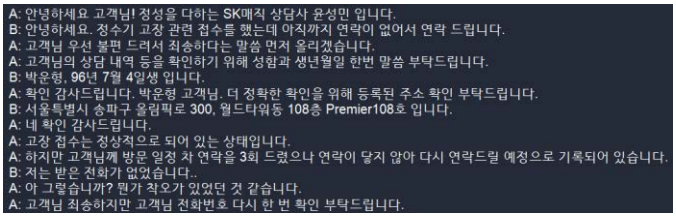


<그림 2> 비식별을 위해 식별된 상태로 비정형 데이터를 정형화하는 모습을 나타낸 흐름도

[그림 2]는 본 연구의 목적에 맞게 수정한 모델을 통과하는 과정을 그린 흐름도이다.

3.1 학습 데이터 형식

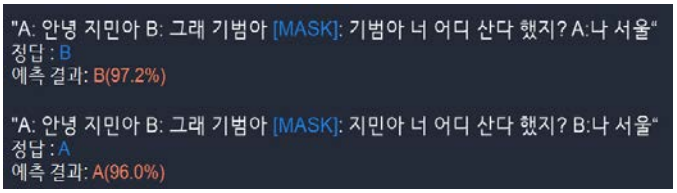
제안한 학습을 진행하고자 [그림 3]과 같은 데이터를 사용하였다. Aihub[5]와 국립국어원 모두의 말뭉치[6]에서 각각 상담 데이터와 SNS 데이터를 전처리하여 학습을 진행하였다. 우리의 목적에 더 적합한 데이터는 상담형 데이터이나, 상담형 데이터는 데이터의 형식이 일정하였기에, 보다 더 어려운 데이터에도 적용 가능한지 실험하기 위해 SNS 데이터를 학습 데이터에 포함시켰다.



<그림 3> Fine-tuning 데이터 예시

3.2 화자 마스킹(Speaker Masking) 기법

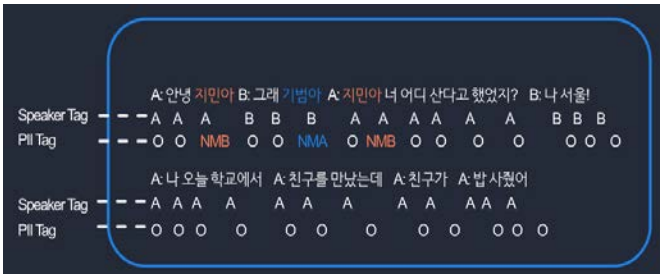
Huggingface[7]에 공개된 KcBERT[8]는 한국어 구어체 데이터를 이용한 사전학습 모델이다. [그림 4]와 같이 우리는 해당 모델에 DialogLM[9] 모델의 사전학습 기법인 Speaker Masking 을 BERT 의 MLM 기법을 활용하여 적용하였다. 이미 사전 학습된 모델에 추가로 학습을 진행하였기에, KcBERT 의 하이퍼 파라미터는 동일하게 설정하였고, 모델의 마스킹 확률을 0.15 에서 0.3 으로 높이고 화자정보를 마스킹하여 학습을 진행하였다. 이러한 과정은 모델이 일련의 연속적인 문자열에서 “A:” 또는 “B:” 이후 문장을 각 화자의 발언으로 인식할 수 있도록 학습하기 위해서이다.



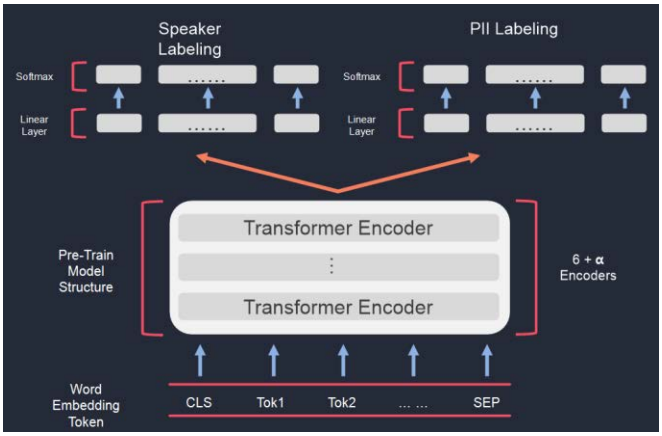
<그림 4> Speaker Masking Format

3.3 다중 라벨 토큰 분류

제안한 모델은 우리의 목적을 해결하기 위해 하나의 어절에 2 개의 태그를 달아 학습을 진행하였다. 하나는 해당 어절의 발언한 사람에 관한 태그이며, 다른 하나는 개인정보의 종류와 누구의 것인지 해당하는지에 관한 태그이다. [그림 5]와 같이 A 의 이름에 관한 태그라면 NMA, NMB 의 이름에 관한 태그라면 NMB, 대화에 참여하지 않은 사람의 이름을 NMO 라고 정하였다. 이러한 분류 작업을 진행하고자 기존 BERT 모델의 Task-Specific 레이어를 [그림 6]과 같이 진행하였다.



<그림 5> Speaker Tag, PII Tag Multi-Label token classification



<그림 6> 본 연구가 제안하는 Language Model 의 구조

4. 실험

다음의 [표 1]은 NMA, NMB, NMO 태그의 정확도이다. 실험의 목적을 위해, 개인정보 중 이름만을 진행하였다. 본 연구에서는 2,000 여개의 대화 데이터로 10 epochs 를 진행하였다.

<표 1> 학습 결과

KoBERT	NMA	NMB	NMO
	0.87	0.91	0.67
KcBERT + Multi-Label Token Classification	NMA	NMB	NMO
	0.89	0.93	0.6
KcBERT + Multi-Label Token Classification + Speaker Masking	NMA	NMB	NMO
	0.9	0.93	0.64

기존의 모델과 비교했을 때, 화자 태그를 추가한 모델은 NMA, NMB 태그에 대해 0.02%의 성능 향상을 보인 반면, NMO 태그는 0.07%만큼 감소하였다. 또한, Speaker Masking 을 추가로 진행한 모델은 앞서 언급한 모델에 비해 전체적으로 향상된 성능을 보인 것으로 확인되었다.

5. 결론

본 연구에서는 기존의 BERT 모델의 추가적인 사전 학습(Speaker Masking)을 진행, Multi-label Token Classification 을 위해 모델의 구조를 수정하였다. 이러한 방식을 통해 기존의 NER 태스크의 한계를 극복하여 대화내에서 식별된 개인정보가 누구의 개인정보인지를 파악할 수 있게끔 하였다.

6. 한계

본 연구의 한계점은 데이터셋의 품질과 양을 첫번째로 꼽을 수 있다. 대화형 데이터 중 SNS 데이터는 오타, 줄임말 등 품질 저하를 일으키는 요인이 된다.

학습에 사용되는 데이터셋의 절대적인 양과 품질 향상을 통해 데이터를 제공한다면 성능이 더 향상될 것으로 예상한다. 제안된 모델은 검증 데이터셋에 존재하는 NMO 태그에 대해 기존 모델보다 낮은 f1-score 를 보이고 있다. 현재까지는 speaker tag 로 인해 NMA, NMB 태그를 식별하는 데에 치우쳐져 있거나, 데이터의 결함이 있는 것으로 분석되고 있다. 향후 이에 대한 정확한 분석이 필요하다.

Acknowledgments:

"This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2021R1F1A1054528) and MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2020-0-01789) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)."

참고문헌

[1]Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova-"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"-2018

[2]Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma,Radu Soricut - "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations"-2019

[3]Yinhan Liu,Myle Ott,Naman Goyal,Jingfei Du,Mandar Joshi,Danqi Chen,Omer Levy,Mike Lewis,Luke Zettlemoyer,Veselin Stoyanov-"RoBERTa: A Robustly Optimized BERT Pretraining Approach"-2019

4]https://github.com/SKTBrain/KoBERT

5]https://aihub.or.kr/

[6]https://corpus.korean.go.kr/

[7]https://huggingface.co/

[8]https://github.com/Beomi/KcBERT

[9]MingZhong,Yang Liu,Yichong Xu,Chenguang Zhu,Michael Zeng - "DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization"-2021