자기지도학습을 이용한 전자의무기록의 개체명 인식

조희수 [1], 이원희 [2]
경희대학교 산업경영공학과
경희대학교 소프트웨어융합학과
{burion002, whlee}@khu.ac.kr

Named Entity Recognition from Electronic Medical Records Using Self-Supervised Learning

Joe Huisu[1], Won Hee Lee[2]
[1]Department of Industrial Engineering, Kyung Hee University
[2]Department of Software Convergence, Kyung Hee University

요 약

전자의무기록(Electronic Medical Records, EMR)을 활용하여 임상 진단, 사망률 예측, 부작용 진단 등 다양한 분야로의 연구가 활발하게 진행되고 있다. EMR에 저장된 많은 양의 환자에 대한 자유 텍스트 데이터는 비정형 데이터로 구성되어 있어 디지털 헬스케어 연구에 활용되는 데 어려움을 겪고 있다. 본 연구에서는 EMR의 비정형 텍스트 기록들로부터 질병, 약물 정보 등과 같은 의학적 단어들을 추출하기 위해 개체명 인식(Named Entity Recognition, NER) 기술을 연결하여 자기지도학습을 통해 모델을 생성하였다. 본 연구에서 제안하는 자기지도학습 기반 개체명 인식 모델은 훈련 데이터와 테스트 데이터를 무작위로 나누었을 때의 당뇨병(diabetes) 등과 같은 중요한 개체명을 인식을 할 수 있었다. 또한 모델의 성능 테스크 결과Macro F1 = 0.82와 Weighted F1 = 0.85의 성능을 달성하였다. 자유 텍스트 임상 기록으로부터 의학적 단어를 인식할 수 있게 함으로써 추후 전자 의료 기록을 통한 특징 추출 및 다양한 디지털 헬스케어 분야의 연구에 이바지할 수 있을 것으로 기대한다.

1. 서 론

전자의무기록(Electronic Medical Records, EMR)은 병원에 내방한 환자에 대한 진료 기록을 기존의 종이에 기록하던 방식에서 컴퓨터를 이용해 전자적 형태로 기록한 데이터들을 의미한다[1]. EMR 안에는 환자에 대한 다양한 종류의 의학 정보가 수집된다. 어떤 질병으로 의심되는 증상, 약물의 투여 지침, 환자의 피드백, 임상의의 위험이나 가정에 대한 의견, 나트륨 평균, 혈당 평균, SAPS ii (Simplified Acute Physiology Score ii) 등이 있다[2]. 최근 몇 년 동안 병원에서 디지털 건강 기록 시스템을 채택하려는 공동의 움직임이 있어 왔다[2]. 예를 들어 미국에서는 기본 디지털 시스템을 갖춘 비연방 급성 진료 병원의 수가 2008년부터 2014년까지 7년 동안 9.4%에서 75.5%로 증가하였다.[3] 가속화되는 고령화와 의학적인 오류로 발생하는 질환들의 증가, 사람의 주관적 견해의 한계로 부딪히는 진단 오류, 치료의 질 개선 요구 등 다양한 의료 산업계 분야의 요구로 인해 디지털 헬스케어 시장 규모는 이미 5.58억에 이르렀다[3]. 다양한 디지털 헬스케어 연구 분야에서 EMR을 활용하여 다양한 연구가 진행되고 있지만, 환자의 개인정보를 포함하고 있는 EMR의 민감한 환자 정보 데이터를 완전히 구조화하고 연구가 가능한 형태로 가공하기까지는 장벽이 높다고 할 수 있다.

따라서 본 연구는 MIMIC-III로부터 획득한 전자 의무 기록 데이터를 사용해 개체명을 인식하고, 인식된 개체명 중 의학 용어 사전에 등록되어 있는 개체명을 연결하여 의학 전문화된 개체명을 인식할 수 있는 모델을 제안한다. NER (Named Entity Recognition)을 진행한 뒤, Context Model로 구축한 의학 언어 사전 CDB(Concept Database)와의 연결 과정을 거치는 해당 모델은 자기지도학습을 통해 학습되었다. 훈련 데이터와 테스트 데이터를 랜덤하게 추출하여 모델의 성능을 평가한 결과 Macro F1 = 0.82, Weighted F1 = 0.85의 성능을 달성하였다. 본 연구에서 제안하는 개체명 인식 모델을 구축함으로써 EMR 데이터 중 환자의 내원 상태에 대한 문자 기록 안에서 특정 질병과의 연관성을 발견하고 진단을 도와주도록 하여, 임상 의사 결정 시스템 구축에 이바지하는 것으로 기대한다. 이를 통해 의료를 개선할 수 있는 디지털 헬스케어 연구를 잠재적으로 활성화시킬 수 있을 것이다.

본 연구는 과학기술정보통신부의 재원으로 한국연구재단의 지원(NRF-2021R1C1C1009436)과 보건복지부의 재원으로 한국보건산업진흥원의 보건의료기술연구개발사업 지원에 의하여 이루어진 것임(HI22C0108).

2. 연구방법

2.1. MIMIC-III

본 연구를 위해 MIMIC-III에서 제공해주는 EMR (Electronic Medical Records) 데이터를 사용하였다[5]. MIMIC-III(Medical

Information Mart for Intensive Care III)는 2001년과 2012년 사이에 Beth Israel Deaconess Medical Center의 중환자실에 입원한 4만 명 이상의 환자에 대하여 비식별화 과정이 완료된 건강 관련 데이터이다. MIMIC-III 데이터 파일은 중환자 정보 시스템의 의료 서비스 제공자가 작성한 문서화된 진행 기록, 병원 전자 건강 기록 데이터베이스의 지속적인 의료 기록, 사회보장국에서 제공하는 병원 외 사망 기록으로부터 구성되었다.

2.2 MedCAT

MedCAT(Medical Concept Annotation Toolkit)은 자유 형식의 건강 관련 기록 텍스트에서 정확하게 질병이나 약물과 같은 의학 정보를 추출하는 NER (Named Entity Recognition) + Linking 라이브러리이다[5]. EMR의 비정형 구조 데이터인 연령, 지역, 성별 등의 속성을 질병과 연결하기 위해 환자에 대한 문자 기록 중에 질병과 관련된 구조화된 데이터가 필요하다. 질병뿐만 아니라 약물, 증상, 절차 또는 기타 의학 데이터에 관한 정보는 자유 텍스트에서만 언급된다. 따라서 MedCAT을 이용하여 환자에 관한 자유 텍스트로부터 새로운 정형화된 EMR 데이터를 구성할 수 있었다.

2.3. Vocabulary and Concept Database

NER + L 모델을 구축하기 위해 MedCAT의 두 가지 파이프라인을 사용하였다[5]. 하나는 Vocab이고, 다른 하나는 CDB(Concept DB)이다. Vocab은 수기로 기록된 환자의 텍스트 데이터에서 철자를 검사하고 단어 임베딩 과정을 진행한다. 가능한 모든 단어 목록을 검사하는 과정이기에 의학 용어랑 무관하게 진행된다. 특정 주제에 해당하는 연구일수록 임베딩도 함께 진행하는 것이 권장되기 때문에[6] 본 연구에서는 BERT(Bidirectional Encoder Representations from Transformers) 모델을 사용하여 임베딩을 진행하였다. 이후 의학적인 연결과 키워드 탐지를 위해 CDB를 구축하였다. 이 데이터베이스는 의료 정보를 추출하기 위해 생물의학 사전에 접속하여 구축한다. 생의학 자원인 UMLS(Unified Medical Language System)과 SNOMED CT(Systematized Nomenclature Of Medicine Clinical Terms)을 활용하여 의료 정보를 추출하였다.

2.4. NER + L (Named Entity Recognition + Linking)

NER + L 모델 생성을 위해 자연어 처리를 사용하여 생물 의학 및 과학 텍스트를 분석하도록 설계된 오픈 소스 ScispaCy를 사용하였다[7]. NER을 통해 개체 후보를 탐지하기 위해 expanded window를 움직이는 방식을 채택했다. 초기에 주어진 문서에 대하여 창 크기를 1, 워드 위치를 0으로 설정해 주었다. 이후 CDB 개념 사전에 존재하는 단어이거나 더 긴 개념의 하위 문자열이라면 윈도우 길이를 1 만큼 확장해 가는 것을 반복하여 최종적으로 전체 길이의 키워드를 추출하도록 하였다. Train epoch는 10을 사용하고, learning rate로는 4.47e-05을 사용하여 학습하였다. 총 160000개의 데이터중 128000 개의 데이터는 train에 사용하고 32000개의 데이터는 test에 사용하였다. 추출된 단어 중 우리가 구성한 의학적 키워드 데이터와 연결하는 Linking 작업을 한다. CDB에서 고유한 이름을 가진 키워드는 UMLS에서도 고유하도록 구성되어 있다. UMLS의 개념 중 95%는 적어도 하나의 고유 이름을 가지고 있으니, 탐지한 키워드 중 각각의 의학 전문 용어로 연결되는 것들을 2차로 context model을 통해 추출하였다.

2.5. 개체명 인식 모델 성능 평가

환자에 대한 자유 텍스트 기록에서 모든 개체들을 추출하고, 감지된 각 단어들이 올바른 UMLS에 연결되어 있는지 확인하였다. NER에서 평가는 토큰이 아닌 엔티티(Entity)별로 수행된다. 이를 위해 Keras에서 제공하는 Callback 클래스 객체를 직접 선언하고 확장해서 훈련 파라미터로 전달하였다. 이후 epoch가 끝날 때마다 Precision과 Recall을 계산한 뒤, 산술 평균인 macro F1-score와 조화 평균인 weighted F1-score 값을 계산하여 성능을 평가하였다. Epoch의 단계가 끝날 때마다 성능 지표 점수를 구하는 것은 seqeval 패키지를 활용하였다.

3. 연구결과

3.1 개체명 인식 결과

임상 기록 텍스트에 대한 키워드 분류 작업 이후 CDB와 연결되는 단어만 거르고 난 뒤의 Entity들을 모두 조사하여 시각화하면 다음 [그림 1]과 같다. 파랗게 표시된 단어들이 모두 의학과 관련된 개체로 인식된 단어들이다. MedCAT 라이브러리에서 제시하는 평가 지표인 Vconcept 점수를 기준으로 색깔의 진함 정도가 결정된다. Vconcept는 인식된 개체명의 의학적 연관성을 나타낸다.[5]

HX: On the day of presentation, this 72 y/o RHM suddenly developed generalized weakness and lightheadedness , and could not rise from a chair. Four hours later he experienced sudden left hand numbness lasting two hours. There were no other associated symptoms except for the generalized weakness and lightheadedness . He denied vertigo .

[그림 1] 환자의 임상 기록 ENR 결과 예시

문장에서 발견되는 Entity는 아래의 [표 1]과 같은 특징을 가진다. Pretty Name은 다양하게 표현된 개체들의 일관된 이름이다. 예를 들어 'LAE', 'left atrial hypertrophy' 는 모두 CDB에 있는 공식 명칭인 'Left atrial hypertrophy'로 표시된다. Context Similarity는 MedCAT에서 제시한 성능 지표인 Vconcept를 사용한다. ICD-10(International Classification of Diseases-10) 질병 코드는 질병에 관한 국제적인 분류 표준 기준이다. CUI(Concept Unique Identifier)는 CDB에서 개체들을 고유하게 분류하기 위한 ID 값이다. Detected Name은 텍스트 중 어떤 단어가 탐지되었는지를 나타내고, 뒤이어서 시작 인덱스와 종료 인덱스로 그 위치를 보여 준다.

[표 1] 탐지된 개체의 특징

Pretty Name	CDB에 연결된 정제된 의학 전문 용어
Context similarity	의학적 진단명과의 유사성 (MedCAT의 Vconcept 점수)
ICD-10 Code	질병 코드
CUI	CDB와 연결되는 고유 코드
Detected Name	탐지된 키워드
Start Index	시작 인덱스
End Index	종료 인덱스

3.2. 개체명 인식 모델 성능 평가 결과

[표 2]는 32000개의 test 데이터 세트에 대해 Confusion Matrix를 구성하고 Precision, Recall을 구한 뒤, 산술 평균인 macro F1-score와 조화 평균인 weighted F1-score 값을 계산한 결과를 보여 준다. Precision은 0.87의 값을 가졌고

Recall은 0.75의 값을 기록했다. Precision과 Recall 값을 사용해 구한 macro F1 점수는 0.82, weighted F1 점수는 0.85로 신뢰할 만한 성능을 보였다.

[표 2] 개체명 인식 모델의 성능 평가 결과

Precision	0.87
Recall	0.75
macro F1	0.82
weighted F1	0.85

4. 결 론

본 연구에서는 EMR에서 환자의 다양한 의학적 정보를 포함하고 있는 비정형화 자유 텍스트 기록을 NER + L 모델을 통해 개체명을 인식을 할 수 있는 방법을 제안하였다. EMR에 저장된 많은 양의 환자에 대한 정보를 포함하고 있으며 비정형화 구조를 가지고 있어 처리가 어렵다. 본 연구에서 제안하고 있는 개체명 인식 기술을 통해 정형화된 데이터를 생성할 수 있었으며, 추후 이를 이용하여 사망률 예측, 질병 예측, 부작용 탐색 등의 다양한 디지털 헬스케어 연구에 활용할 수 있을 것이다. 특정한 질병을 투병 중인 환자들의 EMR 기록에서 나오는 Entity들의 특징을 분석하여 질병에 대한 사전 예방에도 이바지할 수 있다. 수치 데이터로 구성된 의학적 검사와는 다르게 EMR에는 메스꺼움, 불안 증세와 같은 환자의 주관적 경험이나 증상에 대한 기록도 포함되어 있다. 따라서 신체적 검사 결과로 나오지 않는 부분에 대한 탐지를 가능하게 한다.

그러나 본 연구의 주요 한계는 개념 데이터베이스인 CDB의 품질에 크게 의존한다는 것이다. 현재 사용한 UMLS는 대규모 생체 의학 개념 데이터가 존재하지만, 이 안에는 수많은 동의어와 및 줄임말이 존재한다. 특히 의학 전문 분야에서는 약어를 지칭하는 경우가 많기 때문에 혼선을 가져올 수 있다. 예를 들어 'OD'는 과용량인 'overdose'와 매일 한 번인 'once daily"의 두 가지 뜻에 모두 연결될 수 있다. 만약 환자의 진단 상태가 과용량이었을 경우 환자는 약물을 중단해야 하고, 매일 한 번일 경우 주기적인 약물 투여가 필요하다. 즉 매우 다른 뜻을 가지는 같은 단어이기 때문에 실생활의 의학 지식으로 활용된다면 위험을 초래할 가능성이 있다. 새로운 맥락이 기본 개념 임베딩을 업데이트시키면 기존 임베딩의 성능을 저하시킬 수 있기 때문에, 자세한 주석이 사용된 기록을 지정하거나 혼란을 줄 수 있는 단어의 분류가 필요하다. 의학 분야 특성상 약어 사용이 빈번하고, 이로 인한 의사소통의 실수가 발생할 수밖에 없다. 따라서 추후에는 매우 짧은 기록인 의료 기록 안에서도 맥락을 파악해 약어가 어떤 뜻을 의미하는지에 대한 학습을 진행하거나, EMR 기록 중 구조화된 환자의 다른 특성들을 통해 자유 텍스트 안에서 언급되는 약어의 뜻을 유추하는 연구를 진행할 수 있다.

참 고 문 헌

[1] Youngl-Ju Jeun. EMR System and Patient Medical Information Protection. The Korean Journal of Health Service management Vol.7 No.3, 213-224 (2013).

[2] 박영택, 국내의료기관의 전자의무기록시스템 현황 및 발전방향, 정책동향 11권 2호 (2017).

[3] Charles, D., King, J., Patel, V. & Furukawa, M. Adoption of Electronic Health record Systems among U.S. Non-federal Acute Care Hospitals. ONC Data Brief No. 9, 1-9 (2013).

[4] Kang Yoon Lee, Junhewk Kim. Artificial Intelligence Technology Trends and IBM Watson References int the Medical Field. Korean Medinal Education Review 18(2) 51-57 (2016).

[5] Zeljko Kraljevic, Thomas Searle, Anthony Shek. Multi-domain Clinical Natural Language Processing with MedCAT: the Medical Concept Annotation Toolkit. (2010).

[6] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (Database issue) (2004) D267-70. doi:10.1093/nar/gkh061.

[7] M. Neumann, D. King, I. Beltagy, W. Ammar, ScispaCy: Fast and robust models for biomedical natural language processing (Feb. 2019). arXiv: 1902.07669.