자기 지도 학습을 이용한 전자 의무 기록의 개체명 인식

조희수[1, 2] , 이원희[2]
Huisu Joe[1] , Won Hee Lee[2]

[1]Department of industrial and management system engineering, Kyung Hee University, Republic of Korea
[2]Department of Software Convergence, Kyung Hee University, Republic of Korea
{burion002, whlee}@khu.ac.kr

KYUNGHEE UNIVERSITY

Introduction

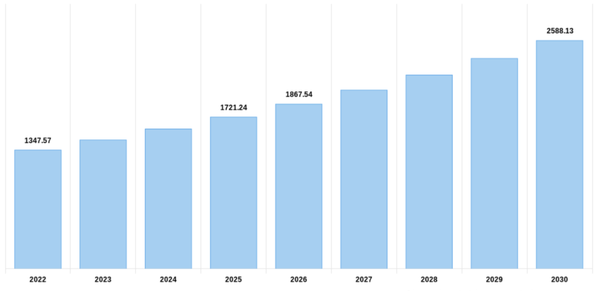

Fig. 1. MarketsAndMarkets에서 발표한 임상 의사결정 시스템 시장 추이

- 전자의무기록(Electronic Medical Records, EMR)은 병원에 내방한 환자에 대한 진료 기록을 종이에 작성하던 방식에서 컴퓨터를 이용해 전자적 형태로 기록한 데이터임.
- 디지털 헬스케어에 대한 시장 규모가 지속적으로 성장함에도 불구하고, EMR에 저장된 많은 양의 환자에 대한 텍스트 데이터는 정리되지 않은 형태로 구성되어 있어 연구에 활용되는 데 어려움을 겪고 있음.
- 본 연구는 EMR의 비정형 문자열 기록들로부터 질병, 약물 정보 등과 같은 의학적 단어들을 추출하기 위해 개체명 인식(Named Entity Recognition, NER) 기술을 연결하여 자기 지도 학습을 통해 모델을 생성하였음.
- 줄글로 되어 있는 텍스트 데이터로부터 의학적 단어를 인식할 수 있게 함으로써 추후 전자 의료 기록을 통한 특징 추출 및 다양한 의학적 의사결정 지원 시스템 연구에 이바지할 수 있을 것으로 기대함.

Dataset and Modeling

Mimic-III Data and MedCAT Library

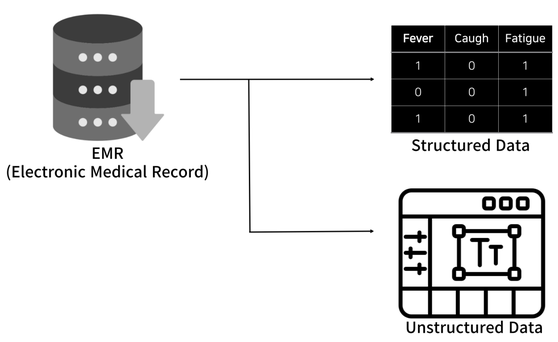

Fig. 2. MIMIC-III 데이터 구조

- **MIMIC-III Database**
  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 익명화 과정이 완료된 건강 관련 데이터임. MIMIC-III 데이터 파일은 중환자 정보 시스템의 의료진들이 작성한 문서화된 환자들의 경과, 병원 전자 건강 기록 데이터베이스의 지속적인 의료 기록, 사회보장국에서 제공하는 병원 외 사망 기록을 합하여 구성되어 있음.

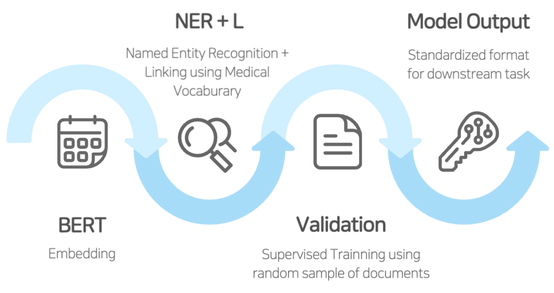

Fig. 3. MedCAT Library Pipeline

- **MedCAT Library**
  MedCAT(Medical Concept Annotation Toolkit)은 자유 형식의 건강 진료 텍스트 기록에서 정확하게 질병이나 약물과 같은 의학 정보를 추출하는 NER (Named Entity Recognition) + L (Linking) Library임. 환자 진료 기록에 대한 구조적이고 정형화된 데이터 추출에 도움을 줌.

Vocabulary and Concept Database

C0006142	Primary Name	Malignant neoplasm of breast
C1134719	Semantic Type	T191: Neoplastic Process
C0035078	Definition	Symptoms of breast cancer may include a lump in the breast, a change in size or shape of the...
C0262926	Synonyms	Breast cancer; Breast ca; Cancer of Breast; Malignant Tumor of Breast; CA; ...
C1217996	Related to	C1458155, C0684503, C0281267, ...
C3366979	Ontologies	SNOMED, MSH, MEDCIN, ...

Fig. 4. 생물의학데이터베이스(UMLS) 예시

- **Concept Database**
  의학적인 연결과 키워드 탐지를 위해 Vocab과 CDB(Concept Database)를 구축하였음. 이 데이터베이스는 의료 정보를 추출하기 위해 생물의학 사전에 접속하여 생성함. 생의학 자원인 UMLS(Unified Medical Language System)과 SNOMED CT(Systematized Nomenclature Of Medicine Clinical Terms)을 활용하여 의료 정보를 추출하였음.

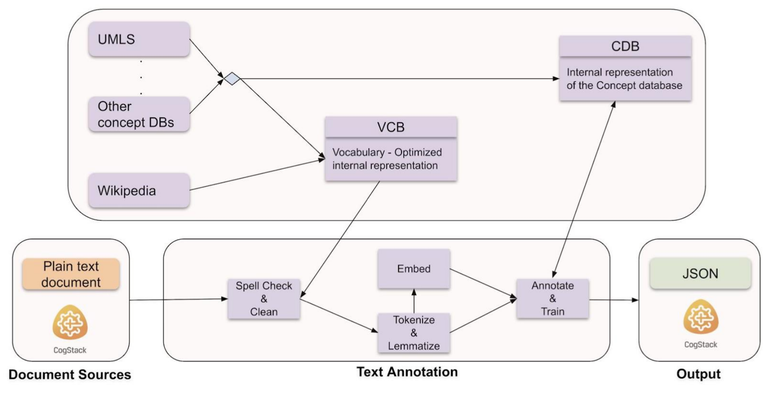

Fig. 5. High Level Representation of Model

- **Linking Preparation**
  추후 줄글 텍스트로부터 추출된 단어 중 우리가 구성한 의학적 키워드 데이터와 연결하는 Linking 작업을 진행. 이 작업을 통해 의학적 전문성을 갖추게 되고, 이를 위해 의학적 사전을 구축하는 것임. 우리가 만든 CDB에서 고유한 이름을 가진 키워드는 생의학 사전인 UMLS에서도 고유하도록 구성되어 있음. UMLS의 개념 중 95%는 적어도 하나의 고유 이름을 가지고 있음.

NER+L (Named Entity Recognition + Linking)

HISTORY OF PRESENT ILLNESS **C0262512** : She is a very pleasant 59-year-old nurse with a history of **C0262926** breast cancer **C0678222** . She was initially diagnosed in June 1994. Her previous treatments included Zometa **C0939788** , Faslodex **C0701491** , and Aromasin **C0876723** . She was found **C0150312** to have disease progression **C0242656** first noted by rising tumor **C0027651** markers

Fig. 6. NER + L 탐지 결과 예시

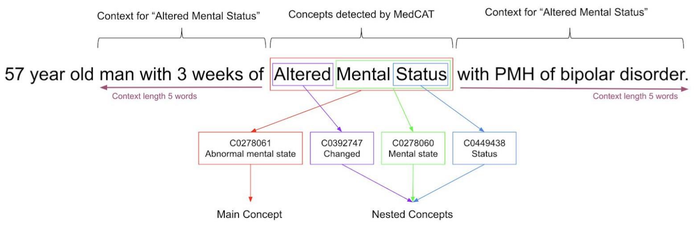

Fig. 7. Biomedical NER + L with nested entities

- **Named Entity Recognition**
  NER+L 모델 생성을 위해 자연어 처리를 사용하여 생물 의학 및 과학 텍스트를 분석하도록 설계된 오픈소스 ScispaCy를 사용하였음. 개체 후보를 탐지하기 위해 expanded window를 움직이는 방식을 채택함. 주어진 문서에 대하여 창 크기를 1, 워드 위치를 0으로 설정해 주었음. 이후 CDB 개념 사전에 존재하는 단어이거나 추출해야 할 단어가 더 길다면 윈도우 길이를 1 만큼 확장해 가는 것을 반복하여 최종적으로 전체 길이의 키워드를 추출하도록 하였음.

- **Linking**
  추출된 단어 중 우리가 구성한 의학적 키워드 데이터와 연결하는 Linking 작업을 진행했음. 탐지한 키워드 중 각각의 의학 전문 용어로 연결되는 것들을 앞서 구성한 CDB를 기준으로 하여 2차로 context model을 통해 추출하였음.

Results

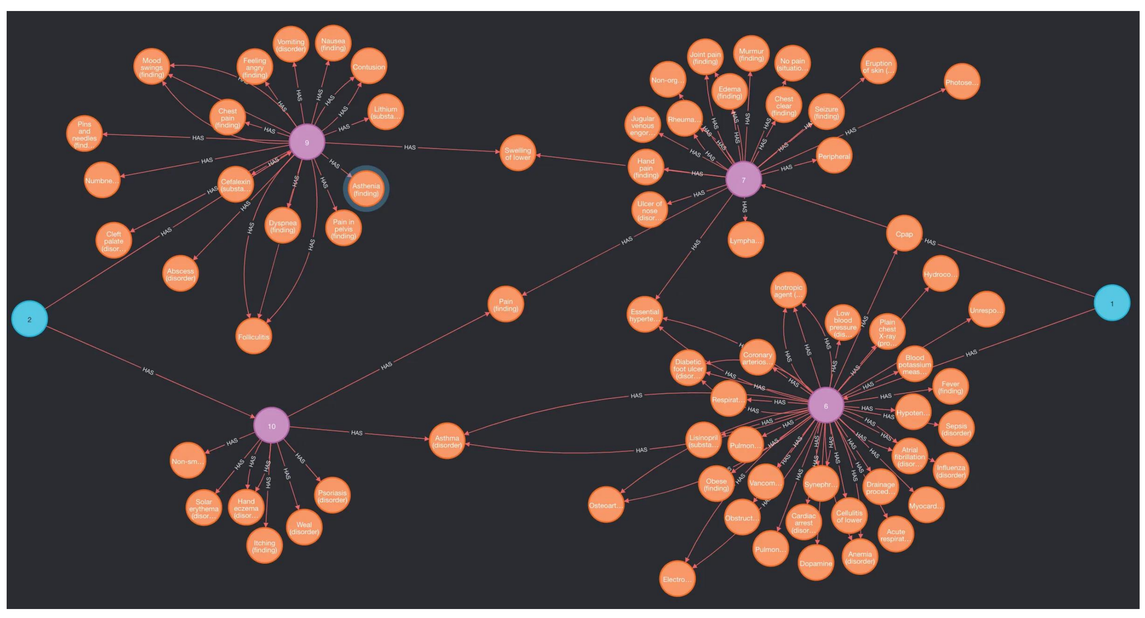

Fig. 8. 문서(분홍색)을 통해 두 환자(파란색)와 관련된 질병, 증상, 약물 및 절차

제시된 그림은 각각 두 개의 문서를 가진 두 환자의 NER + L 결과를 pyvis와 Ipysigma library로 시각화한 것임. 현재 사용한 UMLS는 대규모 생체 의학 개념 데이터가 존재하지만, 이 안에는 수많은 동의어와 및 줄임말이 존재함. 의학 전문 분야에서는 약어를 지칭하는 경우가 많기 때문에 혼선을 가져올 수 있음. 예를 들어 'OD'는 과용량인 'overdose'와 매일 한 번인 'once daily'의 두 가지 뜻에 모두 연결될 수 있음. 따라서 의료 기록 안에서도 맥락을 파악해 약어가 어떤 뜻을 의미하는지에 대한 연구 등이 선행되어야 할 필요성이 있음.

Results

Precision	**0.87**
Recall	**0.75**
Macro F1	**0.82**
Weighted F1	**0.85**

Fig. 9. 모델 성능 평가 결과

생성한 Entity Recognition 모델에 대해 32000개의 test 데이터 세트로 성능을 평가함. Precision은 0.87의 값을 가졌고 Recall은 0.75의 값을 기록함. Precision과 Recall 값을 사용해 구한 macro F1 점수는 0.82, weighted F1 점수는 0.85로 신뢰할 만한 성능을 보였음. 추후 이를 이용하여 사망률 예측, 질병 예측, 부작용 탐색 등의 다양한 디지털 헬스케어 연구에 활용할 수 있을 것으로 기대함.

Conclusions

- 전자의무기록(Electronic Medical Records, EMR)을 활용하여 임상 진단, 사망률 예측, 부작용 진단 등 다양한 분야로의 연구가 활발하게 진행되고 있음.
- 해당 연구에서는 EMR에서 환자의 다양한 의학적 정보를 포함하고 있는 정제되지 않은 자유 텍스트 기록을 NER + L 모델을 통해 개체명을 인식을 할 수 있는 방법을 제안함.
- MIMIC-III EMR 데이터에서 추출한 Entities를 UMLS와 SNOMED 데이터베이스로부터 구축한 Concept Database와 연결하여 의학적인 키워드만 선별하는 과정을 거쳐 전문성을 확보하고자 하였음.
- 자유 텍스트 병원 기록으로부터 의학적 단어를 인식할 수 있게 함으로써 추후 전자 의료 기록을 통한 특징 추출 및 다양한 디지털 헬스 케어 분야의 연구에 이바지할 수 있을 것으로 기대함.

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