보건행정학회지 2020;30(1):15-25 | ISSN 1225-4266
Health Policy and Management Vol.30 No.1, 15-25
https://doi.org/10.4332/KJHPA.2020.30.1.15

REVIEW ARTICLE

보건의료 빅데이터에서의 자연어처리기법 적용방안 연구: 단어임베딩 방법을 중심으로

김한상[1] · 정여진[2]

[1]건강보험심사평가원 심사평가연구실, [2]국민대학교 데이터사이언스학과

A Study on the Application of Natural Language Processing in Health Care Big Data: Focusing on Word Embedding Methods

Hansang Kim[1], Yeojin Chung[2]

[1]Review and Assessment Research Department, Health Insurance Review & Assessment Service, Wonju; [2]Department of Data Science, Kookmin University, Seoul, Korea

While healthcare data sets include extensive information about patients, many researchers have limitations in analyzing them due to their intrinsic characteristics such as heterogeneity, longitudinal irregularity, and noise. In particular, since the majority of medical history information is recorded in text codes, the use of such information has been limited due to the high dimensionality of explanatory variables. To address this problem, recent studies applied word embedding techniques, originally developed for natural language processing, and derived positive results in terms of dimensional reduction and accuracy of the prediction model. This paper reviews the deep learning-based natural language processing techniques (word embedding) and summarizes research cases that have used those techniques in the health care field. Then we finally propose a research framework for applying deep learning-based natural language process in the analysis of domestic health insurance data.

Keywords: Health care big data; High dimensionality; Deep learning; Natural language processing; Word embedding; Word2vec

서 론

건강보험 데이터(claims data)는 전 국민의 의료이용[1] 자료로 환자의 외래방문 및 입원명세서 건 단위로 데이터베이스가 구축되어 있다. 이는 연구목적에 따라 특정 시점별 진료내역 및 환자, 의료기관 단위 등의 다양한 관점으로 분석 가능하여, 정책 시행 및 보완, 결정 등을 위한 근거자료로서 활용되고 있다. 더욱이 4차 산업혁명이 이슈로 대두되면서 보건의료 데이터의 활용방안에 대한 여러 논의가 활발하게 이루어지고 있고[1,2], 2017년 제2차 미래보건의료포럼에서는 보건의료 빅데이터, 인공지능(artificial intelligence) 기술 등의 활성화 방안 등이 논의되었다.

한편, 축적된 정보 등을 이용해 의료이용패턴을 모델링하고, 이를 질병 및 이용행태 예측에 활용하는 것은 보건의료분야의 주요 관심과제 중 하나이다[3]. 최근 보건의료 빅데이터의 축적 및 관련 기술의 발전과 함께 관련 연구가 지속적으로 이루어져 왔고[4-9], 이에 보건의료 빅데이터를 기반으로 정확도 높은 예측모형 개발을 위한 주요 과제들도 논의되었다. 논의의 주된 내용으로는 환자의 과거 건강상태와 시간에 따른 상태변화에 대한 정보가 모델에 반영되어야 하고

1) 건강보험 급여범위 내

Correspondence to: Yeojin Chung
Graduate School of Business Administration, Kookmin University, 77 Jeongneung-ro, Seongbuk-gu, Seoul 02707, Korea
Tel: +82-2-910-5614, Fax: +82-2-910-4332 , E-mail: ychung@kookmin.ac.kr
Received: September 17, 2019, **Revised**: November 20, 2019, **Accepted after revision**: December 10, 2019

[10,11], 의료정보의 복잡성, 고차원, 시간의존성, 불규칙성 등을 고려하여 충분한 설명변수(feature selection)가 필요하다는 것이다[3,10]. 그러나 선형성, 간결성을 강조하여 데이터를 모델에 맞추는(transforming data to model) 전통적 모델링 방법으로는 이런 내용을 충분하게 반영하기 어렵다는 한계가 있었다[12-15]. 이러한 이유로 Goldstein 등[10]은 기존 예측모델연구들의 대부분이 방대한 의료이용 정보에 비해 소수의 가용 가능한 정보만이 활용되었고, 시간에 따른 환자상태 등의 변화에 대한 정보(longitudinal factors)가 제한적으로 활용된 것을 지적하였다.

국내 건강보험 데이터를 활용한 국내 연구환경에서도 과거 정보의 반영 및 충분한 변수선택에 대한 제한점은 동일하게 나타난다. 그 이유로 주요 정보들이 비수치형(범주형) 형태라는 것과 정보범주의 고차원(high dimensional) 등을 들 수 있다. 국내 건강보험 데이터에서 환자의 건강상태 및 의료이용행태 등을 나타내는 대표적 정보로는 진단질환과 처치 및 시술, 검사, 처방 및 조제 약제내역 등을 고려할 수 있다. 해당 정보들은 모두 범주형 형태인 문자코드로 이루어져 있고, 각 정보의 범주가 최소 600개 이상[2]으로 이루어져 있다. 일반적으로 범주형 변수를 분석에 반영하기 위해서는 범주의 개수만큼 0과 1로 이루어진 가변수를 만드는데[3], 이를 반영할 시 데이터의 고차원화로 인한 문제(curse of dimensionality)가 발생된다[16]. 이는 차원이 증가할수록 모델추정에 필요한 데이터의 개수가 기하급수적으로 증가하고, 통계적 유의성이 떨어지게 되어 모델의 성능을 크게 저하시키는 문제를 발생시킨다. 또한 가변수들은 각 범주가 독립적(orthogonal)임을 내포하기 때문에 범주 간의 유사-반대관계 등에 대한 정보를 표현할 수 없어 이런 문제를 더욱 부각시킨다[17].

최근 이 문제에 대해 다른 연구에서는 딥러닝(deep learning)을 이용한 자연어처리(natural language processing)기법을 적용하여 해결방안을 모색하였다. 이는 뉴럴네트워크(neural network)를 활용한 단어임베딩(word embedding) 기법으로 이를 활용하면 범주형 변수의 벡터화를 통해 데이터의 차원축소가 가능하고 문자값들 간의 유사도 비교가 가능해진다. Choi 등[18]은 심부전 예측모델에 Nagata 등[19]은 2형당뇨 예측모델에 뉴럴단어임베딩(neural word embedding) 기법을 이용하여 질환코드와 처치코드, 약제코드 내역을 벡터화하여 분석에 이용하였다. 이 밖에도 다양한 관점에서 이를 활용한 연구들이 지속적으로 이루어지고 있다[20-24].

이러한 배경하에 본 연구의 목적은 다음과 같다. 먼저 국내 보건의료분야에서 아직 생소할 수 있는 단어임베딩기법을 소개하고 의료이용 데이터에 이를 활용한 국외 연구들을 정리하고자 한다. 다음으로 이에 기초하여 국내 건강보험 데이터에 단어임베딩기법을 활용할 수 있는 다양한 관점들과 이를 적용하는 프레임워크를 제시하고자 한다. 마지막으로 건강보험 데이터 분석에 있어 자연어처리기술의 활용 측면에서의 향후 과제에 대해 논의하고자 한다. 이는 보건의료 빅데이터에 대한 활용 논의가 활발한 현 시점에서 딥러닝기술의 활용에 있어 주요 기초 가이드 자료가 될 것으로 기대된다.

뉴럴네트워크 기반 단어임베딩

1. 자연어처리와 단어임베딩

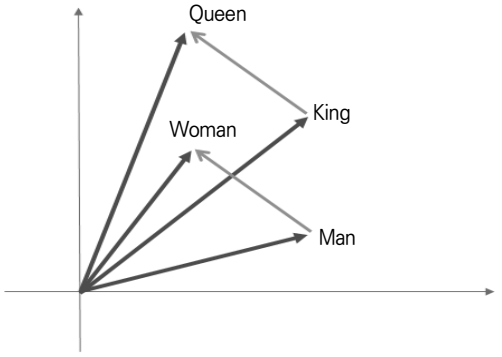
자연어처리는 인간의 언어를 컴퓨터가 처리할 수 있게 해주는 총체적인 기술을 의미하고, 정보검색, 질의응답시스템, 번역 및 통역, 문서작성, 요약분류 등의 여러 분야에서 활용되고 있다. 최근 인공지능영역에서 딥러닝(심층뉴럴네트워크) 기술이 발전함에 따라 자연어처리분야에서도 이를 이용한 관련 연구가 활발히 이루어지고 있다[25].

자연어처리분야에서 단어임베딩은 주요 기술 중 하나로 인식되고 있다[26-28]. 단어임베딩은 주변 단어들의 분포에 기반하여 단어들의 유사도를 계산하고 이를 n차원 벡터로 매핑 시키는 기법이다. 이는 '같은 맥락에서 사용되고 발생되는 단어는 유사한 의미를 나타내는 경향이 있다'라는 언어학의 'distributional hypothesis'에 기초한다[29]. 예를 들어 '왕은 A를 명령했다,' '여왕은 A를 명령했다'라는 문장을 학습하게 되면, 왕과 여왕은 비슷한 맥락에서 사용되었으므로 비슷한 의미를 가지는 것으로 분석되고 벡터공간 내 가까운 위치로 나타내진다. 이런 단어임베딩 방법은 일반적인 원핫인코딩기법을 적용했을 때보다 적은 차원으로 단어의 벡터화가 가능하고, 단어 벡터 간의 수학적 연산을 통해 단어들 간의 유사도를 계산할 수 있다(Figure 1) [30,31].

뉴럴네트워크를 이용한 단어임베딩은 2003년 Bengio 등[32] 제안한 neural network based language model에서 처음 소개되었다. 이는 recurrent neural network language modeling으로 발전되었고[33], 이후 속도 등의 문제를 개선한 word2vec 방법이 개발되어 여러 분야에서 활발하게 활용되고 있다[34]. 최근 word2vec의 문제점을 개선한 단어임베딩 방법이 제안되었는데, GloVe는 주변 단어뿐 아니라 전체 말뭉치(copus)의 전체적인 통계정보를 반영하였고[35], 패스트텍스트(Fasttext)은 skip-gram 모델을 기반으로 '부분단어(subword)'까지를 모델에 반영하였다[36].

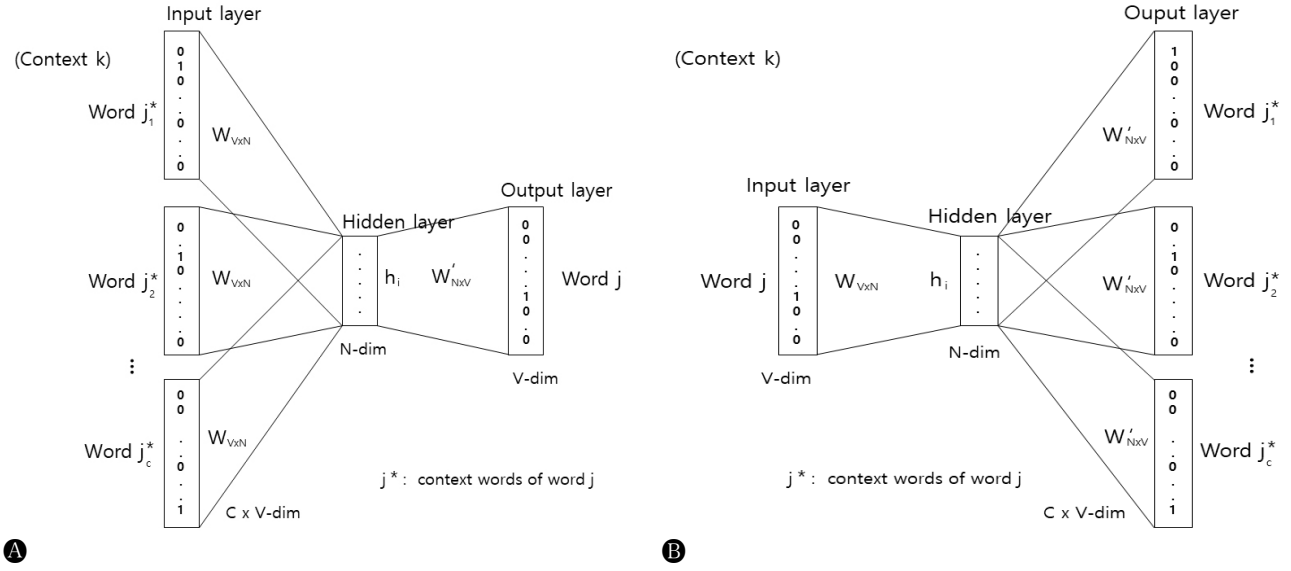
2) 제7차 한국표준질병·사인분류 3단 코드기준
3) 원핫인코딩(one-hot-encoding) 또는 더미변수방법

Word	Encoding
Woman	[1, 0, 0, 0]
Queen	[0, 1, 0, 0]
Man	[0, 0, 1, 0]
King	[0, 0, 0, 1]



Ⓐ Ⓑ

Figure 1. The difference between one-hot_endocding and word embedding [31]. (A) One-hot-encoding. (B) Word embedding.



Ⓐ Ⓑ

Figure 2. CBOW (A) and skip-gram (B) model [37]. CBOW, continuous bag of words.

본 연구에서는 이 방법들 중 가장 대표적인 단어임베딩 방법이면서 토픽분석, 추천시스템 등 다양한 분야에 응용되는 word2vec을 중심으로 건강보험 데이터의 적용방법에 대해 서술한다.

2. Word2vec

Word2vec은 구글에서 2013년에 발표한 단어임베딩모형이다. 이 모형은 학습방식에 따라 “continuous bag of words (CBOW)”와 “skip-gram” 두 가지 모델로 나뉜다. CBOW 모델은 주변 단어들로 그 사이의 중간 단어를 예측하는 모델이고, skip-gram 모델은 반대로 하나의 단어를 통해 주변 단어들을 예측하는 모델이다. 두 모델 모두 학습을 통해 단어별로 n차원의 히든레이어(hidden layer) 값이 추정되는데, 이 값이 해당 단어의 벡터값이 된다(Figure 2) [37].

CBOW 모델은 예측하고자 하는 ‘단어(target word)’와 동일 ‘문장’ 안에서의 ‘주변단어(window)’에 기반하여 학습된다. 학습과정은 문장마다 구성단어들이 순차적으로 하나의 출력변수(target)가 되고, 선택된 단어의 주변단어들이 입력변수(input) 값이 되는 뉴럴네트워크모형에 기반하여 손실함수[4]를 최소화하는 매개변수(parameter) 값을 계산하게 된다. Skip-gram 모델은 CBOW 모델과 구조는 비슷하지만 입력값과 예측하고자 하는 방향이 반대로 구성된다(Figure 2) [37].

Word2vec 방법 적용 시 학습할 데이터(문장 리스트)와 주변단어의

4) 실제값과 예측값의 차이에 대한 크기를 나타내는 함수

범위(window size), 적용모델방법(CBOW, skip-gram), 벡터의 차원 등의 정의가 필요한데, 이는 시뮬레이션을 통한 모델의 성능비교를 통해 적정값을 찾게 된다.

3. 단어임베딩의 건강보험 데이터 적용

건강보험 데이터에 단어임베딩기법을 적용하는 기본 아이디어는 '단어'를 진단상병 및 모든 진료내역 코드들로 정의하는 것이다. 즉 환자가 의료기관에 방문하여 발생된 진단상병, 처치내역, 약제 등의 코드를 각각 하나의 단어로 정의하고, 이런 단어들이 다양한 정의에 따라 하나로 묶여 문장을 만든다. 이는 위에서 언급한 '같은 맥락에서 사용되고 발생되는 단어는 유사한 의미를 나타내는 경향이 있다'는 'distributional hypothesis'를 확장하여 '환자의 의료이용내역[5]들을 문장으로 정의했을 때, 비슷한 주변코드들과 자주 발생되는 코드는 유사한 의료상황에서 나타날 확률이 높다'는 가정에 기반하여 각 코드 간 벡터값의 위치를 결정한다. 따라서 환자의 의료이용내역 코드들로 어떻게 문장을 구성했는지에 따라(주변단어의 구성), 임베딩벡터 값은 다양하게 나타난다.

문장구성 외에도 하나의 문장 내에서 각 코드가 다른 코드들에게 영향을 미치는 범위(window size), 임베딩모델 방법(CBOW, skip-gram), 각 코드들을 나타내는 임베딩변수[6]의 개수(n)를 결정하고, 뉴럴네트워크 학습을 통해 각 코드별 n개의 임베딩변수 값이 산출된다(Figure 3).

4. 연구사례

서론에 언급한 것과 같이 최근 보건의료데이터에 자연어처리기법을 활용하는 연구가 활발히 이루어지고 있다. 이 중 진단질환, 약제, 치료내역 코드 등을 활용한 사례도 꾸준하게 소개되고 있는데, 본 장에서는 이런 사례들의 연구목적과 이를 위한 단어임베딩기법을 어떻게 활용하였는지에 대해 소개하고자 한다.

2016년 Choi 등[23]은 청구자료(claims dataset)에서 단어임베딩기

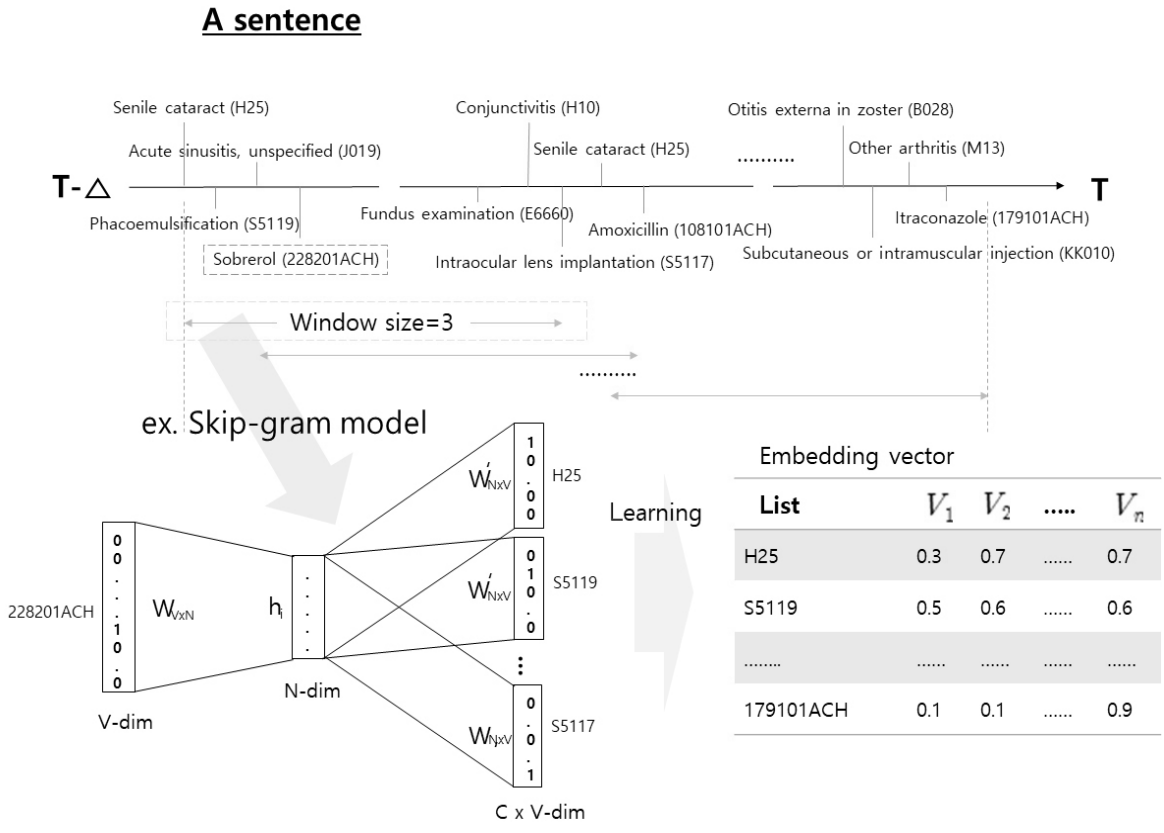


Figure 3. Word embedding in claim data. Codes: 7th Korean Standard Classification of Diseases: health insurance fee schedule (2018. 2): drug benefit list.

5) 이후 진단상병, 처치내역, 약제, 치료재료 등의 내역을 '의료이용내역'으로 명칭함
6) 본 논문에서는 단어임베딩기법을 통해 생성되는 n차원 벡터를 '임베딩변수'로 명칭함

법을 적용하는 방법을 소개하였다[23]. 해당 연구에서는 1년 동안 한 환자에게서 발생되는 행위코드(Current Procedural Terminology), 약제코드(National Drug Code), 국제질병사인분류(International Classification of Diseases-9th revision, ICD-9), 진단검사코드(Logical Observation Identifiers Names and Codes)를 하나의 문장으로 정의하였고, 이를 통해 각 코드별로 임베딩변수 값을 산출하고 각 코드 간의 유사도를 계산하였다.

2016년 Choi 등[18]은 전자건강기록(electronic health record, EHR) 자료를 이용한 심부전 예측모델에서 단어임베딩기법을 적용하였다[18]. 문장은 특정기간 내(약 4년) 한 환자에게서 발생되는 진단상병, 약제코드, 행위코드를 묶어서 정의하였고, 단어임베딩기법을 적용하여 각 코드들을 100개의 임베딩변수로 나타내었다. 그 결과 기존 원핫인코딩 방법을 적용했을 때보다 심부전 발생예측의 정확도가 향상되는 것으로 나타났다.

2017년 Bai 등[21]은 EHR 자료에서 구조화된 정보(ICD-9 code)와 비구조화된 정보(clinical note)를 결합하여 단어임베딩기법을 적용하는 방법(JointSkip-gram)을 소개하였다. 연구에서 문장은 환자의 병원 방문당 발생되는 진단질환 코드들과 임상노트의 단어들로 정의하여 각 코드 간의 연관성과 코드와 임상노트의 단어 간의 연관성을 산출하였다. 임상전문가 평가결과 기존 모형[7]보다 연관성이 더 명확한 것으로 나타났고, 환자의 다음 방문의 진단코드의 예측모형에 이를 임베딩변수로 반영 시 기존 모형들보다 예측의 정확도가 향상되는 것으로 나타났다.

2017년 Che 등[22]은 EHR 자료를 이용하여 단기간 위험예측에 효과적인 모델을 제안하였다. 한 환자에게 발생된 전체 진단질환, 약제내역을 문장으로 정의하고 단어임베딩기법을 적용하여 200개의 임베딩변수로 변환하여 모델에 적용하였다[22]. 그 결과 90, 180일 심부전, 당뇨 조기 예측모델의 정확도가 크게 향상되는 것으로 나타났다.

2018년 Nagata 등[19]은 건강검진(health checkup), 청구(claim)자료를 이용하여 2형 당뇨 발병 예측모델을 개발하였는데, 여기서 청구자료의 진단질환과 약제내역에 단어임베딩기법을 적용하였다[19]. 이 연구에서는 매달 발생되는 청구내역의 진단질환 코드와 약제내역 각각을 하나의 문장으로 정의하였고 이를 200개의 가변수로 변환하였다. 그리고 추가적으로 건강검진변수들까지 포함하여 최종 모델의 설명변수로 이용하였다. 연구결과 임베딩변수를 설명변수로서 모델에 추가했을 때 예측정확도가 높게 나타났다.

2018년 Jin 등[20]은 EHR 자료를 통해 환자의 진단상병 이력을 하나의 문장으로 정의하였고, 여기에 단어임베딩기법을 적용 후 이를 이용한 시계열기반 심부전 예측모형을 제안하였다. 연구결과 기존 방법을 적용한 것보다 단어임베딩을 이용했을 때의 예측의 정확도가 높아지는 것으로 나타났다.

2018년 Zhang 등[24]은 EHR 자료를 통해 환자의 방문에서 발생되

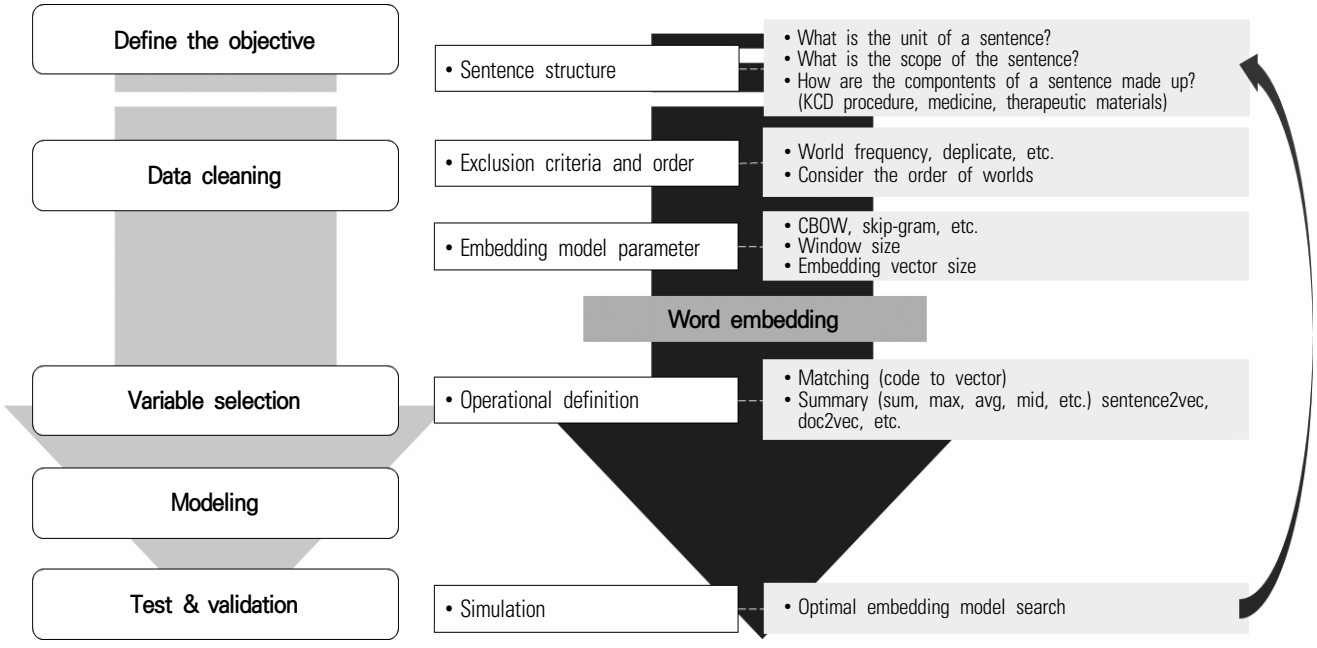


Figure 4. Analytics framework. KCD, Korean Standard Classification of Diseases; CBOW, continuous bag of words.

7) 토픽모형(latent dirichlet allocation)

는 코드들을 하나의 문장으로 정의하였고, 이를 이용하여 환자 단위 시계열 예측모델인 "Patient2Vec"을 제안하였다. 연구에서는 이 모델을 이용하여 불필요한 재입원 예방을 위한 재입원의 조기예측을 하였고, 기존 시계열모형보다 예측정확도가 높게 나타났다.

단어임베딩 적용 연구프레임워크

본 장에서는 앞서 소개한 연구사례들에 기초하여 국내 건강보험데이터 분석 시 고려될 수 있는 다양한 활용관점에 대해 정리하고자 한다. 일반적으로 데이터 분석 프로세스는 분석기획, 데이터전처리, 변수선택, 모델링, 검증 및 테스트의 순으로 이야기할 수 있다. 이 과정에 따라 단어임베딩기법 적용을 위한 고려사항을 순차적으로 정리하면 Figure 4와 같다.

1. 문장의 구조

일반적으로 문장은 환자 단위로 구성된다. 즉 한 문장은 한 환자에 발생된 내역을 기준으로 만들어진다. 세부적으로는 이에 기초하여 다양한 관점으로 나눠질 수 있는데, 선행연구들을 살펴보면 주로 문장의 단위, 범위, 그리고 구성요소 등으로 구분되었다(Table 1).

1) 문장의 단위(unit)

먼저 단기, 중·장기모델이냐에 따라 문장을 환자 단위 의료이용 전체(혹은 일부) 내역 또는 환자의 의료기관 방문당(명세서 건당)으로 정의할 수 있다(Table 1). 연구의 목적이 미래의 질병 및 이용행태 등의 중장기 예측(또는 설명)이라면 환자 단위 과거 의료이용내역 전체(혹은 특정시간 단위)를 하나의 문장으로 정의해야 임베딩변수 값에 중장기 의료정보패턴이 반영될 수 있다. 앞서 소개한 선행연구에서 Choi 등[18]은 약 4년간의 의료이용내역을, Choi 등[23]은 1년간의 의료이용내역을, Nagata 등[19]은 월에 발생된 의료이용내역을 문장을 정의하였다. 만일 단기예측(또는 설명)이 목적일 경우에는 환자의 장기간의 방문 의료이용패턴보다는 관심 시점의 바로 직전 어떤 의료이용을 했는지가 더 중요하므로 방문당(명세서당)을 하나의 문장으로 정의하는 것이 적정하다. Bai 등[21]과 Zhang 등[24]은 환자의 방문당 발생되는 코드들과 임상기록의 단어들로 문장을 정의하였다.

2) 문장의 범위(scope)

건강보험 데이터에서 분석대상군(처리, 대조군)은 일반적으로 특정 연령대, 보험자격 등의 환자 특성, 입원/외래, 방문주기 등의 의료이용행태, 그리고 특정 질환 등으로 정해지지만 임베딩학습을 위한 문장구성 시에는 분석대상보다 넓은 범위의 환자군의 기록을 대상으로 할 수 있다(Table 1). 예를 들어 분석모델의 목적이 심부전 예측이라면 문장의 범위는 심부전이 발생된 환자의 이전 의료이용내역 등이 고려될 수 있다. 다만 심부전이 발생된 환자의 데이터가 충분하지 않다면, 확보된 전체 데이터(전체 환자군)를 문장의 범위로 지정하는 것이 더 좋은 분석결과를 얻을 수 있다. Choi 등[18]은 문장의 범위를 분석대상군만으로[8] 정의했을 때와 전체 환자군으로 정의했을 때 임베딩변수를 도출하고 이를 모델에 반영했을 때의 예측정확도를 비교하였는데, 단어임베딩기법 적용 시에는 전체 환자군을 모두 활용하는 것이 정확도가 높게 나타났다. 따라서 연구의 성격 및 데이터에 따라 문장범위에 대한 다각도적인 고려가 필요하다.

3) 문장의 구성요소(component)

임베딩학습을 위한 문장은 진단상병(주, 부상병), 약제, 행위(검사 포함), 치료재료 코드들의 조합으로 만들어질 수 있는데, 이는 연구의 목적과 분석자료의 고유값의 개수, 분석데이터의 크기 등을 고려하여 연구자가 주관적으로 결정하게 된다(Table 1). 2016년 Choi 등[18]은 약 26만 명의 환자의 4년간 의료이용 데이터를 이용하여 총 38,599

Table 1. Define sentence in claims data

Components	Units	
	Short term (visit)	Mid term (visits) & long term (visit history)
Disease	Bai et al. [21] (2017)*	Jin et al. [20] (2018)
Disease & medicine	-	Che et al. [22] (2017)*, Nagata et al. [19] (2018)
Disease & procedure	-	-
Disease & medicine & procedure	Zhang et al. [24] (2018)	Choi [23] et al. (2016)*, Choi [18] et al. (2016)*

*Apply word embedding using entire patient group instead of a group of patients with specific characteristics (sex, age, etc.), disease, utilization behavior (inpatients/outpatients), etc.

8) 처리군 혹은 처리군과 대조군

개[9]의 진단상병, 약제, 행위코드를 문장의 구성요소로 정의해 단어임베딩기법을 적용하였다. 그리고 Che 등[22]은 약 22만 명의 환자의 14,969,489개의 관측값 기준, 총 14,690개의 진단상병, 약제코드를 문장의 구성요소로 정의하였다.

2. 제외기준 및 문장 내 순서 결정

단어임베딩은 비슷한 주변단어의 분포에 기초하여 벡터값을 산출한다. 따라서 특정 코드의 빈도가 낮게 발생되면 임베딩변수 값에 오차가 발생 확률이 커질 수 있어 분석 시 이에 대한 제외기준이 필요할 수 있다. Bai 등[21]은 연구에서 전체 발생빈도가 임상노트의 단어는 50 미만, 진단질환 코드는 5 미만은 제외하고 임베딩변수 값을 산출하였다. 또한 의료이용의 특성상 특정 환자의 경우 동일한(또는 유사한) 의료이용내역이 단기간에 중복적으로 발생될 수 있는데, 이에 대한 처리방법의 결정이 필요하다. Choi 등[23]은 청구자료에서 짧은 기간에 발생되는 중복코드와 순서는 정보의 가치가 없다고 여겨 기간내[10] 발생되는 중복값을 제거하였다.

또한 문장 내 순서에 대한 고려도 필요한데, 이는 앞서 설명한 것처럼 문장은 환자의 의료이용내역을 기반으로 하고 진단코드, 약제코드, 행위코드 등의 조합으로 구성되기 때문에 코드의 배치를 어떻게 하느냐에 따라 학습에 영향을 줄 수 있기 때문이다. Word2vec 알고리즘은 가까이에 위치해 있는 단어들 간의 유사도를 높게 학습시킴으로써 문장 내에서 맥락을 고려한다. 자연어의 경우 문장 내에서 단어의 순서가 자연스럽게 결정되기 때문에 이를 그대로 사용하는 것이 정당하지만 의료내역의 경우 문장 내에서 코드의 순서가 특정한 의미를 지니지 않는다면(예를 들어 시간의 흐름에 따른 상병코드의 변화가 아닌 동일 명세서에서 발생되는 코드들의 자의적 순서) 데이터 자체의 순서에 따라 word2vec 알고리즘을 학습시키는 것은 바람직하지 않다. 만일 연구목적에 따른 특정 순서가 없다면 무작위 순서를 학습 시 고려할 수 있다. Choi 등[23]은 반복훈련(epoch)마다 코드의 위치를 랜덤하게 섞어(shuffle) 단어임베딩기법을 적용하였다.

3. 단어임베딩모델 파라미터 값 설정

최종 단어임베딩 적용을 위해서는 word2vec 방법 기준 CBOW, skip-gram 모델 선택, 주변단어의 범위설정 그리고 가변수의 개수를 정의해줘야 한다. 먼저 모델 선택의 경우 skip-gram 모델의 성능이 빈도가 낮은 단어에서 좋은 성능을 보이는 것으로 소개되었지만[34], 이와 반대되는 연구결과도 있어[38], 연구 시 두 가지 모델에 대한 성능의 비교는 필요할 것으로 판단된다. 그리고 임베딩변수의 개수는 구글가이드라인에서는 많을수록 좋은 것으로 소개[11]되었는데, 기존연구들에서는 100, 200, 400개 등으로 지정해 주었다. 다만 레코드 수가 충분히 많지 않을 때, 너무 많은 임베딩변수는 분석결과에 오차를 발생시킬 수 있으므로 이를 고려하여 선택할 필요가 있다. 마지막으로 주변단어의 범위는 구글 가이드라인에는 skip-gram의 경우 10 근처, CBOW의 경우 5 근처로 권고하고 있다. 기존 연구에서는 작게는 2부터 많게는 20까지 적용하였고, 여러 값들의 예측정확도를 고려하여 최종 범위를 결정하였다.

4. 단어임베딩 벡터의 조작적 정의

단어임베딩을 통해 각 진료내역 코드들은 n개의 임베딩변수로 표현되고 이는 건강보험 데이터의 코드와 매칭(matching)하여 사용되나(Figure 3), 최종 분석데이터의 레코드[12] 단위(record units)에 따라 임베딩변수의 조작적 정의가 필요하게 된다. 데이터 레코드의 단위가 한 환자라면, 분석기간 내 각 환자의 모든 방문(명세서)당[13] 발생되는 진료내역 코드들의 정보들을 하나의 정보로 표현해야 모델의 설명변수로 반영할 수 있다(Table 2). 예를 들어 Jin 등[20]은 환자의 방문단위 데이터에 기반한 시계열(종단면)분석을 수행하였는데, 진단상병 코드의 임베딩변수를 조작적 정의 없이 분석에 반영하였다(Table 2). 이는 진단상병(주상병)의 경우 환자의 방문당 하나의 코드가 발생되어 분석데이터에 바로 매칭하여 사용할 수 있었기 때문이다. 반면, Choi 등[18]은 진단상병, 약제코드, 행위코드의 임베딩변수 값을 이용하여 환자 단위 횡단면분석을 수행하고자 하였다. 따라서 각 진료내역 코드별로 산출된 임베딩변수를 조작적 정의를 통해 환자단위의 값으로의 변환이 필요한데, 연구에서는 이를 환자별로 합산하는 방법을 적용하였다(Table 2).

한편, 임베딩변수의 조작적 정의방법은 합계 외에도 다양하게 있을 수 있다. 최근 word2vec 방법에 기초한 다양한 방법들이 활발하게 논의되고 있는데, 대표적으로 doc2vec (document to vector), sentence2vec (sentence to vector) 방법 등이 있다[39,40]. 이는 word2vec을 확장한 개념으로 문장과 문서를 n차원의 벡터로 임베딩하는 방법이다. 여기서의 문장과 문서는 단어들의 모임으로 건강보험 데이터 측면에서는 환자에게서 발생되는 진료코드들의 묶음정보 등이 될 수 있어 분석단위에 따라 다양하게 적용 가능하다.

9) 진단상병 1,460개, 약제코드 17,769개, 행위코드 9,370개
10) 연구에서는 1년을 3등분한 기간을 기준으로 함.
11) https://code.google.com/archive/p/word2vec/
12) 또는 분석데이터의 행(row)
13) 외래, 입원 모두 방문당으로 표시하였고, 이는 건강보험 데이터에서 명세서당을 의미함. 다만, 입원의 경우는 한 에피소드를 의미함

Table 2. Record units and embedding vectors

Record units			Embedding vector		
			V_1	....	V_n
Patient (1)	Operational definition value (per patient)		Value		Value
	Visit (1)	Medical records (1)	Value	....	Value
		....	....	....	....
		Medical records (m_{ij})		....	
		Operational definition value (per visit)	Value	....	Value
	...	Medical records (1)	Value	....	Value
		....	....	....	....
		Medical records (m_{ij})		....	
		Operational definition value (per visit)	Value	....	Value
	Visit (v_i)	Medical records (1)	Value	....	Value
		....	....	....	....
		Medical records (m_{ij})	Value	....	Value
		Operational definition value (per visit)	Value	....	Value
....	....				
Patient (p)	....				

p, number of patients; v_i, number of visits by patient; m_{ij}, number of records by patient visit.

5. 시뮬레이션

앞서 살펴본 바와 같이 문장의 정의부터 임베딩변수의 적용까지는 다양한 접근이 가능하다. 이는 이에 대한 전반적인 검토가 수반되어야 함을 의미한다. 대부분의 기존 연구들은 다각도적인 시뮬레이션을 통해 이를 비교하고 최종 모델을 선정하였다.

6. 적용 예시

본 장에서는 단어임베딩 적용 프레임워크의 이해를 돕기 위해 분석사례를 소개하였지만 문장의 정의부터 임베딩변수의 산출까지만 다루었다. 분석에는 2008-2015 한국의료패널[14] 자료를 이용하였고, python3의 keras_talk 라이브러리를 이용하여 word2vec을 적용하였다.

먼저 문장은 4년간 발생된 환자별 주상병 이력으로 정의하였고, 외래방문내역만 문장범위에 포함하였다. 그리고 주상병코드의 빈도가 5 미만이면 해당 코드는 분석에서 제외하였고, 시간에 따른 중복상병은 분석에 포함하였다. 또한 문장 내 순서는 방문일자에 따라 나열하였다(Table 3). 끝으로 모델은 skip-gram을 사용하였고, 주변단어의 범위를 5로 지정하여 200번 반복하여(iterations) 학습시켰다.

분석결과 분석데이터 내에서 주상병코드의 고유값은 총 903개로 나타났다. 이는 주상병코드를 설명변수로 이용하기 위해 원-핫 인코딩을 사용한다면 약 903개의 가변수가 필요하다는 의미와 같다. 문장수(환자 수)는 19,815개로, 평균 문장당 단어 수(환자당 외래 방문횟수)는 49개로 나타났다. 최종적으로 빈도수가 5 미만인 단어(주상병코드)를 제외하고 771개의 주상병코드에 대한 50개의 임베딩변수를 산출하였다(Table 3).

Table 3. Sentences and word embedding vector

Sentence				
Patient ID				
1	"E14, M15, E14, J00, J00, M15, ... E14, E14"			
...	...			
19,815	"A09, A09, J00, J00, J00, K29, ... H71, J00"			
	↓	↓	↓	
KCD code	V_1	V_2	...	V_{50}
J00	-0.165949	-0.101547	...	-0.018292
E14	-0.498270	0.752446	...	0.177178
...	...	...	...	...
M15	0.915924	0.569913	...	0.012296

KCD, Korean Standard Classification of Diseases.

임베딩변수 값의 적정성을 살펴보기 위해 2형 당뇨상병(E11) 근접한 값을 가지는 상병을 살펴본 결과, '죽상경화증(I70),' '외이염(H60),' '비뇨계통의 기타 증상 및 징후(R39),' '고관절 및 대퇴부위의 근육 및 힘줄의 손상(S76),' '병적 골절을 동반한 골다공증(M80),' '노년백내장(H25)' 등으로 나타났다. 이 중 대부분은 당뇨에 따른 합병증으로 발생될 수 있는 상병 등으로, 단어임베딩의 목적에 따라 당뇨

14) http://www.khp.re.kr/

병 환자에게서 비슷하게 발생되는 코드들이 유사한 값으로 매칭된 것으로 보인다. 다만 임베딩변수에 대한 정확한 판단은 임상적 근거와 각 질환코드별 충분한 데이터가 뒷받침되어야 한다.

결 론

보건의료데이터는 방대한 정보를 가지고 있지만 이질성, 불규칙성(longitudinally irregular), 잡음(noise) 등의 문제로 인해 분석에 있어 제한점도 존재한다[22]. 특히 주요 의료이용내역 등의 정보가 텍스트(코드)로 되어 있는데, 이를 반영할 시 설명변수의 고차원화 등의 문제로 이 정보에 대한 활용이 제한적이었다. 최근 딥러닝기술이 발전됨에 따라 이를 활용하여 이런 문제들을 극복하려는 시도가 지속적으로 이루어지고 있는데, 자연어처리기법 중 하나인 단어임베딩의 적용도 이런 시도 중 하나이다. 해외에서는 이에 대한 연구들이 지속적으로 발표되고 있는데, 대부분의 연구에서 진료내역 코드들을 단어임베딩을 통해 모델의 설명변수로 반영했을 때[15] 기존 방법들보다 예측정확도 및 연관성 등이 높게 나타났다[19,20,22]. 따라서 딥러닝 기반 단어임베딩기법의 적용은 보건의료 빅데이터의 활용범위 확장 및 분석모델의 설명력 향상 등에 기여될 수 있을 것으로 판단된다. 이에 따라 본 연구에서는 딥러닝 기반 단어임베딩과 이를 보건의료분야에서 활용한 연구사례들을 소개하였고, 최종적으로 국내 건강보험데이터에 적용할 수 있는 다양한 관점들과 방법 등을 정리하여 분석 프레임워크를 제안하였다.

본 연구에서는 단어임베딩기법 중 하나인 word2vec을 중심으로 설명하였는데, 본문에 제시된 프레임워크는 건강보험데이터에 단어임베딩기법을 적용하는 전반적인 가이드로 여러 임베딩모델 방법에도 확장 적용이 가능하다. 또한 연구의 목적에 따라 건강보험 데이터에서 발생될 수 있는 연구사례들을 주로 다루었지만, 해외 몇몇 연구에서는 EHR 내 임상전문가의 텍스트 기록의 분석에도 단어임베딩기법을 활용하였다[41-43]. 이에 국내 전자의무기록(electronic medical record)과 의약품 안전사용서비스(drug utilization review)의 예외사유 등의 텍스트 정보가 발생되는 자료에서도 단어임베딩기법은 다양하게 활용될 수 있을 것으로 판단된다.

끝으로, 본 연구는 건강보험데이터를 분석함에 있어 단어임베딩기법의 적용과정 및 방법들에 대한 설명에 초점을 맞췄고, 이해를 돕기 위해 의료패널 데이터를 이용한 간단한 적용사례를 소개하였다. 그러나 향후 단어임베딩기법을 적용한 모델개발 연구 시 의료패널 데이터의 활용뿐만 아니라 분석목적에 따라 정보의 크기 등의 제한점이 존재할 수 있어 건강보험 전체 데이터를 활용한 연구들이 필요할 것으로 판단된다. 그리고 최근 해외에서는 기존 단어임베딩기법을 적용하는 것이 아닌 보건의료데이터의 복잡성[44], 시간불규칙성[45] 등을 반영한 다양한 응용 단어임베딩모델들이 개발되고 있다. 국내에서도 보건의료데이터에 적정한 모델 검토 및 개발에 대한 다양한 관점의 연구 및 논의가 필요하다.

15) 데이터가 충분히 많을 경우

감사의 글

이 논문은 한국연구재단 연구비(NRF-2016R1C1B1010940)와 산림청(한국임업진흥원) 산림과학기술 연구개발사업(2019150B10-1923-0301)에 의해 수행되었다.

ORCID

Hansang Kim: https://orcid.org/0000-0001-7347-7342;
Yeojin Chung: https://orcid.org/0000-0003-4117-2880

REFERENCES

1. Chang E, Kim D, Lee J, Yang B, Hwang J, Kwak S. A study on the advancement of utilization of medical big data. Wonju: Health Insurance Review & Assessment service; 2016.
2. Kang H. National-level use of health care big data and its policy implications. Sejong: Korea Institute for Health and Social Affairs; 2016.
3. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016;6:26094. DOI: https://doi.org/10.1038/srep26094.
4. Himes BE, Dai Y, Kohane IS, Weiss ST, Ramoni MF. Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. J Am Med Inform Assoc 2009;16(3):371-379. DOI: https://doi.org/10.1197/jamia.M2846.
5. Amarasingham R, Moore BJ, Tabak YP, Drazner MH, Clark CA,

Zhang S, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. Med Care 2010;48(11):981-988. DOI: https://doi.org/10.1097/MLR.0b013e3181ef60d9.

6. Saltzman JR, Tabak YP, Hyett BH, Sun X, Travis AC, Johannes RS. A simple risk score accurately predicts in-hospital mortality, length of stay, and cost in acute upper GI bleeding. Gastrointest Endosc 2011;74(6):1215-1224. DOI: https://doi.org/10.1016/j.gie.2011.06.024.
7. Mani S, Chen Y, Elasy T, Clayton W, Denny J. Type 2 diabetes risk forecasting from EMR data using machine learning. AMIA Annu Symp Proc 2012;2012:606-615.
8. Kennedy EH, Wiitala WL, Hayward RA, Sussman JB. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Med Care 2013;51(3):251-258. DOI: https://doi.org/10.1097/MLR.0b013e31827da594.
9. Tabak YP, Sun X, Nunez CM, Johannes RS. Using electronic health record data to develop inpatient mortality predictive model: Acute Laboratory Risk of Mortality Score (ALaRMS). J Am Med Inform Assoc 2014;21(3):455-463. DOI: https://doi.org/10.1136/amiajnl-2013-001790.
10. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017;24(1):198-208. DOI: https://doi.org/10.1093/jamia/ocw042.
11. Pham T, Tran T, Phung D, Venkatesh S. DeepCare: a deep dynamic memory model for predictive medicine. In: Kim J, Shim K, Cao L, Lee JG, Lin X, Moon YS, editors. Advances in knowledge discovery and data mining: 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings. Cham: Springer; 2017. pp. 30-41.
12. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018;19(6):1236-1246. DOI: https://doi.org/10.1093/bib/bbx044.
13. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018;1:18. DOI: https://doi.org/10.1038/s41746-018-0029-1.
14. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: predicting clinical events via recurrent neural networks. JMLR Workshop Conf Proc 2016;56:301-318.
15. Zhang E, Robinson R, Pfahringer B. Deep holistic representation learning from EHR. Proceedings of the 2018 12th International Symposium on Medical Information and Communication Technology (ISMICT); 2018 Mar 25-28; Sydney, Australia. Piscataway (NJ): IEEE; 2018.
16. Bellman, R. Adaptive control processes: a guided tour. Princeton (NJ): Princeton University Press; 1972.
17. Rodriguez P, Bautista MA, Gonzalez J, Escalera S. Beyond one-hot encoding: lower dimensional target embedding. Image Vis Comput 2018;75:21-31. DOI: https://doi.org/10.1016/j.imavis.2018.04.004.
18. Choi E, Schuetz A, Stewart W, Sun J. Medical concept representation learning from electronic health records and its application on heart failure prediction [Internet]. Ithaca (NY): arXiv; 2016 [cited 2019 Sep 15]. Available from: https://arxiv.org/abs/1602.03686v1.
19. Nagata M, Takai K, Yasuda K, Heracleous P, Yoneyama A. Prediction models for risk of type-2 diabetes using health claims. Proceedings of the BioNLP 2018 Workshop; 2018 Jul 18-23; Melbourne, Australia. Stroudsburg (PA): Association for Computational Linguistics; 2018.
20. Jin B, Che C, Liu Z, Zhang S, Yin X, Wei X. Predicting the risk of heart failure with EHR sequential data modeling. IEEE Access 2018;6:9256-9261. DOI: https://doi.org/10.1109/access.2017.2789324.
21. Bai T, Chanda A, Egleston B, Vucetic S. Joint learning of representations of medical concepts and words from EHR data. Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2017 Nov 13-16; Kansas City, USA. Piscataway (NJ): IEEE; 2017.
22. Che Z, Cheng Y, Sun Z, Liu Y. Exploiting convolutional neural network for risk prediction with medical feature embedding [Internet]. Ithaca (NY): arXiv; 2017 [cited 2019 Sep 15]. Available from: https://arxiv.org/abs/1701.07474v1.
23. Choi Y, Chiu CY, Sontag D. Learning low-dimensional representations of medical concepts. AMIA Jt Summits Transl Sci Proc 2016;2016:41-50.
24. Zhang J, Kowsari K, Harrison JH, Lobo JM, Barnes LE. Patient2vec: a personalized interpretable deep representation of the longitudinal electronic health record. IEEE Access 2018;6:65333-65346. DOI: https://doi.org/10.1109/access.2018.2875677.
25. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. IEEE Comput Intel Mag 2018;13(3):55-75. DOI: https://doi.org/10.1109/mci.2018.2840738.
26. Yang Liu, Zhiyuan Liu, Tat-Seng Chua, Maosong Sun. Topical word embeddings. Proceedings of the 29th AAAI Conference on Artificial

Intelligence; 2015 Jan 25-29; Austin, USA. Menlo Park (CA): Association for the Advancement of Artificial Intelligence; 2015.
27. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space [Internet]. Ithaca (NY): arXiv; 2013 [cited 2019 Sep 15]. Available from: https://arxiv.org/abs/1301.3781.
28. Kiela D, Hill F, Clark S. Specializing word embeddings for similarity or relatedness. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing; 2015 Sep 17-21; Lisbon, Portugal. Stroudsburg (PA): Association for Computational Linguistics; 2015.
29. Harris ZS. Distributional structure. Word 1954;10(2-3):146-162. DOI: https://doi.org/10.1080/00437956.1954.11659520.
30. Trask A, Gilmore D, Russell M. Modeling order in neural word embeddings at scale [Internet]. Ithaca (NY): arXiv; 2015 [cited 2019 Sep 15]. Available from: https://arxiv.org/abs/1506.02338.
31. Mikolov T, Yih WT, Zweig G. Linguistic regularities in continuous space word representations. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2013 Jun 9-14; Atlanta, USA. Stroudsburg (PA): Association for Computational Linguistics; 2013. pp. 746-751.
32. Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. J Mach Learn Res 2003;3:1137-1155.
33. Mikolov T, Kombrink S, Burget L, Cernocky J, Khudanpur S. Extensions of recurrent neural network language model. Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2011 May 22-27; Prague, Czech Republic. Piscataway (NJ): IEEE; 2011.
34. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality [Internet]. Ithaca (NY): arXiv; 2013 [cited 2019 Sep 15]. Available from: https://arxiv.org/abs/1310.4546.
35. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014 Oct 25-29; Doha, Qatar. Stroudsburg (PA): Association for Computational Linguistics; 2014.
36. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017;5:135-146. DOI: https://doi.org/10.1162/tacl_a_00051.
37. Rong X. Word2vec parameter learning explained [Internet]. Ithaca (NY): arXiv; 2014 [cited 2020 Mar 24]. Available from: https://arxiv.org/abs/1411.2738.
38. Sahlgren M, Lenci A. The effects of data size and frequency range on distributional semantic models. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing; 2016 Nov 1-4; Austin, USA. Stroudsburg (PA): Association for Computational Linguistics; 2016.
39. Le QV, Mikolov T. Distributed representations of sentences and documents. Proceedings of the 31st International Conference on Machine Learning; 2014 Jun 21-26; Beijing, China. Stroudsburg (PA): International Machine Learning Society; 2014.
40. Giatsoglou M, Vozalis MG, Diamantaras K, Vakali A, Sarigiannidis G, Chatzisavvas KC. Sentiment analysis leveraging emotions and word embeddings. Expert Syst Appl 2017;69:214-224. DOI: https://doi.org/10.1016/j.eswa.2016.10.043.
41. Minarro-Gimenez JA, Marin-Alonso O, Samwald M. Exploring the application of deep learning techniques on medical text corpora. Stud Health Technol Inform 2014;205:584-588.
42. De Vine L, Zuccon G, Koopman B, Sitbon L, Bruza P. Medical semantic similarity with a neural language model. Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM 2014); 2014 Nov 3-7; Shanghai, China. New York (NY): Association for Computing Machinery; 2014.
43. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission [Internet]. Ithaca (NY): arXiv; 2019 [cited 2019 Dec 4]. Available from: https://arxiv.org/abs/1904.05342.
44. Choi E, Bahadori M, Searles E, Coffey C, Sun J. Multi-layer representation learning for medical concepts [Internet]. Ithaca (NY): arXiv; 2016 [cited 2019 Dec 4]. Available from: https://arxiv.org/abs/1602.05568v1.
45. Cai X, Gao J, Ngiam K, Ooi B, Zhang Y, Yuan X. Medical concept embedding with time-aware attention [Internet]. Ithaca (NY): arXiv; 2018 [cited 2019 Dec 4]. Available from: https://arxiv.org/abs/1806.02873.