creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

의학박사 학위논문

이종언어 임상기록 내 흡연 음주 정보 추출 및 분류를 위한 자연어처리 알고리즘 개발

Natural Language Processing Algorithm for Extraction and Classification of Smoking and Drinking Information from Bilingual Clinical Notes

2023년 8월

서울대학교 대학원

의학과 협동과정 의료정보학 전공

배 예 슬

이중언어 임상기록 내 흡연 음주 정보 추출 및 분류를 위한 자연어처리 알고리즘 개발

지도교수 윤 형 진

이 논문을 의학박사 학위논문으로 제출함

2023년 4월

서울대학교 대학원

의학과 협동과정 의료정보학 전공

배 예 슬

배예슬의 의학박사 학위논문을 인준함

2023년 7월

위 원 장 최 진 욱 (인)

부위원장 윤 형 진 (인)

위 원 지 의 규 (인)

위 원 박 상 민 (인)

위 원 고 태 훈 (인)

초 록

이중언어 임상기록 내 흡연 음주 정보 추출 및 분류를 위한 자연어처리 알고리즘 개발

서울대학교 의과대학 의학과

협동과정 의료정보학 전공

배예슬

흡연과 음주는 중요한 질병 위험 인자이자, 의학 연구에 필수적인 중요변수이다. 따라서 환자의 현재 흡연과 음주 상태에 대한 정보를 확보하는 것이 중요하나, 우리나라는 전자의무기록 작성 시 한국어와 영어가 혼용되어 사용되므로 자유진술문으로 구성된 임상기록에서 이를 자동으로 확보하는 연구는 드물다. 또한 의료 영역에서 한국어가 포함된 대규모 corpus가 부재하므로 self-supervised learning인 transformer 방법론을 적용하여 fine-tuning하기 어렵다. 이에 본 논문에서는 Shifted Positive Pointwise Mutual Information (SPPMI)와 transformer 기반의 서로 다른 두 방법론으로 구조화 되어있지 않은 한영 혼용 임상기록에서 키워드를 추출하고 문서를 분류하는 연구를

진행하였다.

흡연 연구에서는 총 4,711개의 임상기록이 활용되었으며 python package인 Soynlp를 사용하여 두문자어를 치환하고, 임상기록을 일반화하였다. SPPMI를 사용하여 기록 내 각 단어를 벡터화 하였고, cosine similarity를 계산하여 흡연 상태를 분류하였다. word co-occurrence, pointwise mutual information (PMI), normalized pointwise mutual information (NPMI)와 같은 다른 방법과 비교하여, 본 연구는 키워드 추출 정확도(macro F1-score)를 20% 이상 향상시켰다. 또한 문서 내 흡연 관련 키워드를 추출하는데 제안한 SPPMI 방법이 가장 높은 precision을 보였으며 흡연 상태를 분류하는데 91.49%의 높은 정확도를 보였다.

음주 관련 정보를 추출하고 문서를 분류하기 위하여 총 4,996개의 임상기록이 활용되었으며 multilingual transformer를 fine-tuning 하였다. 제안한 cross-lingual language model - robustly optimized BERT pretraining approach (XLM-RoBERTa) 모델은 다른 방법과 비교하여 높은 키워드 추출 정확도(macro F1-score 84.7%)를 보였으며, 78.23%의 precision과 95.33%의 recall을 보여 우수한 성능을 보였다. 또한 술의 종류와 알코올 소비패턴의 보조정보를 추가로 학습한 경우, 음주 정보에 대해 문서를 분류하는 성능이 multilingual BERT는 9.87%, XLM-RoBERTa는 11.7%의 향상되었다.

본 연구는 한영 혼용으로 작성된 자유진술문 구조의 임상 기록에서 유의미한 정보를 추출하고 문서를 분류한 최초의 연구이다. 또한 현재 활발히 사용되고 있는 transformer에 한영 혼용 임상 기록을 적용하여 유의미한 정보를 확보하는 최초의 연구이기도 하다. 향후 흡연 음주

정보뿐 아니라 임상기록 내 다양한 정보를 확보하는 데 본 연구에서 제안하는 방법을 활용할 수 있을 것이다. 향후 다양한 진료과와 의료기관에서 생성된 임상기록을 대상으로 유효성 검증 연구가 필요할 것이다.

주요어 : 자연어처리, 흡연, 음주, 전자건강기록, 이종언어처리, 트랜스포머

학 번 : 2021-31273

목 차

제 1 장 서 론 1

제 1 절 연구 배경 1

제 2 절 연구 목적 4

제 2 장 SPPMI를 이용한 흡연 키워드 추출 및 상태분류 7

제 1 절 연구 방법 7

제 2 절 결과 14

제 3 장 Transformer를 이용한 음주 키워드 추출 및 상태분류 27

제 1 절 연구 방법 27

제 2 절 결과 35

제 4 장 종합 고찰 45

제 5 장 결 론 54

참고문헌 55

Abstract 61

표 목차

Table 1. Overall summary of clinical notes data..... 9

Table 2. Examples of seed keywords 11

Table 3. Comparison of precision between SPPMI-based keyword expansion and six baseline models 15

Table 4. Examples of extracted keywords 17

Table 5-1. Comparison of overall smoking status classification accuracy 19

Table 5-2. Comparison of overall smoking status classification precision..... 20

Table 5-3. Comparison of overall smoking status classification recall..... 21

Table 6-1. Comparison of never smoker classification accuracy 23

Table 6-2. Comparison of past smoker classification accuracy 24

Table 6-3. Comparison of current smoker classification accuracy 25

Table 7. Summary of basic patient statistics..... 28

Table 8. Overview of the annotated dataset 35

Table 9. Overall comparison of the extraction performance of different NLP models 37
Table 10. The extraction performance with respect to different languages used in the dataset 38
Table 11. Examples of the extracted alcohol-related keywords 39
Table 12. General statistics of current status in the dataset 40
Table 13. General statistics of amount status in the dataset 40
Table 14. General statistics of binge status in the dataset 41
Table 15. Pattern extraction performance in current status 43
Table 16. Pattern extraction performance in amount status 43
Table 17. Pattern extraction performance in binge status 43

그림 목차

Figure 1. Overall summary of our SPPMI-based keyword expansion method..... 10
Figure 2. A number of extracted keywords with respect to their frequencies 26
Figure 3. A few examples symbols and acronyms preprocessed during the symbol replacement and acronym expansion 30
Figure 4. An overview of the transformer-based language model for extracting patients' alcohol-related information from their free-text clinical notes 32
Figure 5. A summary of classification model; Multilingual + additional information..... 34

제 1 장 서론

제 1 절 연구 배경

생활습관이 사망이나 질병에 미치는 영향에 대한 연구는 1960년 대 Doll 과 Hill[1]의 흡연에 대한 연구를 시작으로 현재까지 꾸준히 이어지고 있다. 통계청에 따르면 2021년 우리나라 19세 이상 성인의 흡연율①은 18.2%[2]로 2016년 22.6% 이후 지속적인 감소 추세를 보이고 있다. 음주의 경우, 고위험음주율 ② 이 12.4%[3]에 그치는데 반해, 월간 폭음률③은 32.4%[4]로 19세 이상 성인 3명 중 1명은 폭음을 하고 있다.

직접 담배 사용은 기대 수명을 감소시키고 사망 위험을 증가시키며, 모든 암 중 약 30%가 직접 담배 사용과 관련이 있고, 흡연량이 많고 흡연 기간이 길수록 질병 발생 및 사망 위험이 높다[5,6]. 질병관리청에 따르면, 우리나라에서 2019년 흡연으로 인한 사망자수는 만 30세 이상 남성 약 5만 1천명, 여성 약 7천명으로 총 6만명에 달한다[6].

음주는 질병부담의 중요 기여요인으로, 소량의 음주라도 고혈압[7], 암[8]을 유발할 수 있다. 특히, 과도한 음주는 신경정신장애[9], 심뇌혈관질환[10], 간경변[11], 암[12] 등의 중요 위험요인이며 영양 결핍[13]과 수면장애[14]를 유발하고, 우울감[15] 및 자살률 증가[16]와도 관련이 있다.

---

① 현재흡연율: 평생 일반담배 5갑(100개비) 이상 피웠고, 현재 일반담배(궐련)를 피우는 분율 (국민건강영양조사, 질병관리청)

② 고위험음주율: 1회 평균 음주량이 남자의 경우 7잔 이상, 여자의 경우 5잔이상이며, 주 2회 이상 음주하는 분율 (국민건강영양조사, 질병관리청)

③ 최근 1년 동안 월 1회 이상 한번의 술자리에서 남자의 경우 7잔(또는 맥주 5캔 이상), 여자의 경우 5잔(또는 맥주 3캔 이상) 음주한 분율 (국민건강영양조사, 질병관리청)

또한 교통사고, 폭력, 자살 등의 치명적 사고를 유발할 수 있다[17]. 또한 음주 문제는 고위험군 또는 취약집단에 더욱 큰 영향을 미친다[18].

다행히 흡연과 음주는 교정가능한 위험인자 (modifiable risk factors)[19]이다. 현재 흡연자이거나 고위험 음주자일지라도 주치의의 적절한 중재를 통해 건강한 생활습관을 함양하여 일차 및 이차 예방을 이룩할 수 있다. 이에 진료 현장에서 환자의 흡연과 음주 정보를 확보하는 것이 매우 중요하다.

흡연, 과도한 음주가 개인뿐 아니라 사회에 미치는 부담은 매우 크지만, 동시에 개선이 가능한 건강위험행동이기 때문에 의료진이 적극적으로 개입하여야 할 필요가 있는 건강요인으로 볼 수 있다. 이를 위해서는 정확한 정보 확보가 필수적이나 현재는 대부분 전자의무기록 (Electronic Medical Record, EMR)으로부터 수동으로 추출하거나, 설문지를 통해 획득하는 등 많은 비용과 인력을 사용해서 정보를 획득하고 있는 실정이다. 이러한 방식은 많은 비용과 인력이 소모된다는 단점이 있다.

임상의가 임상소견을 전자의무기록에 기록할 때, 통제 어휘 또는 자유진술문 (free text)을 이용하게 된다. 자유진술문이란 특별한 규정 없이 일상적 언어를 사용하여 표현한 문장을 말하는데[20], 생활습관 행태는 임상의에 의하여 주로 자유진술문 형태로 기록된다. 자유진술문으로 기록된 문건에서 의미 있는 정보를 확보하는 대표적 방법으로 자연어처리 (natural language processing, NLP)[21] 기법을 들 수 있다. 자연어처리란 자연어의 의미를 분석하여 컴퓨터가 처리할 수 있도록 하는 일련의 과정[21]으로, 현재 자연어처리를 이용하여 자유진술문 형태의 임상기록에서 약 처방, 문제 목록, 종합적인

임상정보를 추출하는 시도가 진행되고 있다[22]. 텍스트 처리 기술은 단어를 벡터 공간에 표현하는 단어 임베딩 (word embedding)[23-25] 방법이 등장한 이후 급격히 발전하였으며, 딥러닝 모델이 자연어처리에 이용되는 결정적인 계기가 되었다. 신경망 언어 모형은 단어의 의미를 주변 단어와의 관계를 통해 벡터로 표현하는 기술을 적용해 의료 용어들의 의미론적 유사성 구조 표현에 활용한다.

임상기록 (clinical note)은 대부분 의사가 환자를 진료하는 동시에 생성되기 때문에, 비구조적 (unstructured)인 경우가 많고 다양한 임상 정보를 포함하게 된다. 임상기록에서 특정 정보를 추출하거나 문서를 분류 할 때, 기존에는 UMLS (unified medical language system)[26]과 같은 개념 정규화 (concept normalization) 기반 리소스나 cTAKES[27]와 같은 자연어처리 toolkit이 종종 활용되었다[28-30].

Jonnagaddala 등[31]은 영문 임상기록에서 흡연 정보를 추출하여 구조화된 형식으로 변환하였다. Shoenbill 등[32]은 오픈 소스 자연어처리 도구인 cTAKES를 사용하여 생활습관에 대한 평가 및 조언에 대한 정보를 추출하였다. Wang 등[33]은 자연어처리를 기반으로 알코올, 약물, 니코틴 사용과 같은 물질 사용 (substance use) 정보를 추출하였다. 추출 시 언어 모델 (language model)과 의료 특화 정보 (domain knowledge)을 함께 적용하였을 때 가장 좋은 성능을 보여주었다. Lix 등[34]은 Canadian Primary Care Sentinel Surveillance Network에서 환자의 알코올 사용을 자동으로 분류하는 지원 벡터 머신 분류기를 제안하였다. Alzoubi 등[35]은 미리 정의된 알코올 관련 키워드셋을 사용하여 알코올 관련 문장을 탐지하였는데, unigram과 bigram을 포함한 다양한 머신러닝 분류기를 사용하여 현재 알코올을 섭취하는

사람을 분류하였다. 이와 유사하게, Afshar 등[36]은 term-frequency, inverse document-frequency (TF-IDF)를 사용하여 트라우마 센터에 내원한 환자들 중 알코올을 남용하는 경우를 식별하였다. To 등[37]은 자연어처리 기반 분류기를 비외상 환자를 대상으로 적용하고 일반화 가능성을 살펴보았다. Fraile-Navarro 등[38]은 알코올 사용 패턴을 5가지로 분류하기 위해, 수동 주석 (manual annotation)을 사용하였다.

그러나 대부분의 선행연구는 영어로 기록된 임상기록을 기반으로 모델을 구축하였다. 따라서 선행 연구의 방법론은 비영어권 국가의 임상기록을 활용한 실험에서는 제약이 있을 수 있다. 또한, 임상기록 내 흡연 또는 음주 관련 문장이 이미 전처리된 데이터 리소스로 제공되거나 추출할 수 있도록 한 경우가 대부분이므로, 한국어 기반의 의료 기록 데이터 리소스가 부재한 우리나라의 연구에는 적용하기 어려운 한계가 있다.

제 2 절 연구 목적

선행연구를 살펴보면, 대부분의 기존 자연어처리 모델이 영어 또는 단일언어를 기반으로, 의료 특화 데이터 셋이 아닌 일반목적의 데이터 셋을 사용하여 구축되었음을 알 수 있다. 우리나라의 경우 임상기록의 자연어처리가 어려운데, 몇 가지 요인이 있다. 먼저, 한국어와 영어가 혼용되어 데이터 전처리와 학습에 어려움이 있다. 또한 미국의 경우와 같이 문장 (sentence)으로 기록하기 보다는 구 (phrase)로 간략히 표현하여 기록한 경우가 많으며, 작성자 편의에 맞는 약어를 사용하는 빈도가 높다. 다음으로, 흡연, 음주 정보는 비정형 자유형식 텍스트로

기록되는 경우가 많아 표준화 및 규격화된 데이터 관리가 어렵다. 더불어 임상기록 내 방대한 텍스트 중, 흡연과 음주와 관련된 텍스트는 일부분일 뿐이다. 이에 용어 표준화가 이루어지지 않은 실정에서 규칙 기반의 자연어처리 기술이 적용되기 어렵다. 더불어 의학 분야에서는 한영 혼용 임상기록 기반의 트레이닝을 위한 대규모 데이터셋의 부재로 자연어처리 연구가 거의 이루어지지 않은 실정이다.

즉, 의료 영역에서 한국어가 포함된 대규모 corpus가 부재하므로 self-supervised learning을 적용한 transformer 방법론을 사용하기 어려우며 이에 본 논문에서는 shifted positive pointwise mutual information (SPPMI)와 transformer 기반의 서로 다른 두 방법론으로 구조화 되어있지 않은 한영 혼용 임상기록에서 키워드를 추출하고 문서를 분류하는 연구를 진행하고자 한다.

제 2장에서는 unsupervised learning 방법 중 하나인 SPPMI를 활용해, 방대한 데이터 기반의 선행 훈련 과정 없이 임상기록 내 흡연 관련 키워드를 추출하고 문서를 분류할 수 있는 방법을 제안하고자 한다.

제 3장에서는 bidirectional encoder representations from transformer (BERT)와 같은 transformer 기반 방법론을 활용하여 음주 관련 키워드를 추출하고 문서 분류를 진행하고자 한다. 선행 훈련 과정을 위하여 음주 관련 키워드를 수동으로 임상기록 내에서 추출하고 문서에 주석 (annotation)을 붙이는 단계를 거쳤다.

본 연구는 연구윤리지침을 준수하였다. 또한 익명화된 의무기록 정보를 후향적으로 분석하는 것으로 설계되었기 때문에 서울대학교병원 연구윤리검토위원회는 연구 대상자의 서면 동의에 대한 면제와 함께 연구를 승인하였다. (IRB No. H-1906-076-1040, H-2207-084-

1340)

제 2 장 SPPMI를 이용한 흡연 키워드 추출 및 상태분류

본 장에서는 한국어와 영어가 혼용된 비구조적인 임상기록에서 흡연 관련 키워드를 추출하고 문서를 분류하고자 한다. 전자의무기록 내 임상기록은 대부분 방대한 임상 정보를 담고 있기 때문에 원하는 정보의 양은 문서 내 아주 일부일 뿐이다. 본 장에서 제안하는 방법은, 흡연 상태와 관련 있는 키워드를 추리고 식별하며, 이를 기반으로 문서 분류까지 진행한다. 또한 제안하는 모델은 unsupervised method인 SPPMI를 활용하였으므로 의료 분야에 특화된 대규모 트레이닝 데이터셋을 필요로 하지 않는다. 따라서 이는 비영어권 국가의 전자의무기록을 활용하는 하나의 실용적인 방법일 수 있다.

제 1 절 연구 방법

1) 데이터 확보

2017년 1월 1일부터 12월 31일까지 서울대학교병원 가정의학과 및 호흡기내과에 내원한 당뇨병 또는 만성폐쇄성폐질환 환자의 외래초진 또는 외래경과 기록을 추출하여 실험에 활용하였다. 데이터 추출에는 서울대학교병원 clinical data warehouse 인 SUPREME (Seoul National University Hospital Patients Research EnvironMEnt)을 사용하였다. 중복되는 기록을 제거하고, 최종 연구에는 가정의학과

의무기록은 총 3,512개, 호흡기내과 의무기록은 총 1,199개가 사용되었다.

각 임상기록에는 의사가 기록한 환자의 전반적인 병력 (주호소, 내원사유, 통증유무, 주관적 소견, 신체검진 소견, 사회력, 과거력, 추정진단, 치료방침 등)과 진찰소견이 포함되어 있었다. 평균 157.04개의 token이 있었으며 token 길이의 범위는 1에서 589까지 다양하였다. 임상기록은 자유진술문로 작성되었기 때문에 동일한 용어나 개념이라도 의사마다 표현하는 방식이 다양하였다. 또한 영어와 한국어 단어가 모두 포함되어 있어 일반적인 키워드 추출이 복잡한 특성이 있었다.

2) 흡연 상태의 정의 및 데이터 전처리

환자의 흡연 상태에 따라, 각 임상기록을 '현재 흡연자 (current smoker)' , '과거 흡연자 (past smoker)', '비흡연자 (never smoker)', '정보 없음 (absence of information)'의 네 카테고리로 분류하였다. 현재 흡연자는 현재 흡연을 하고 있는 경우, 과거 흡연자는 흡연을 하였으나 현재 끊은 경우, 비흡연자는 평생 흡연을 한 적이 없는 경우, 알 수 없음은 임상기록 내 흡연 관련 정보가 없는 경우이다.

흡연 상태에 대한 주석 작업은 3단계로 진행하였다. 먼저, 3명의 annotator가 1차 labeling을 하고, 그 다음 간호사 경력이 있는 1명의 annotator가 문서 전수를 2차 검수한 후 마지막으로 1인의 가정의학과 전문의가 문서 전수를 3차 검수하였다. 본 과정에서는 흡연 상태에 대한 annotation만 진행하였으며, 문서 내 흡연 관련 키워드를 수동 주석 (manual annotation)처리하지는 않았다. Table 1에서 흡연 상태에

따른 임상기록의 분포를 알 수 있다.

Table 1. Overall summary of clinical notes data.

	Family medicine	Pulmonary and critical care medicine	Total
Current smokers	1,046	84	1,130
Past smokers	547	431	978
Never smokers	399	144	543
Absence of information	1,520	540	2,060
Total	3,512	1,199	4,711

3) Shifted Positive Pointwise Mutual Information (SPPMI)-based keyword extraction

본 실험은 SPPMI을 사용하였다. SPPMI는 종단간 비지도학습 방법(end to end unsupervised method)으로, 주석이 달린 데이터나 model training이 필요치 않다. 이에 의학 영역에서, 많은 시간과 비용이 소요되는 수동 주석작업 없이 적용가능한 방법 중 하나이다.

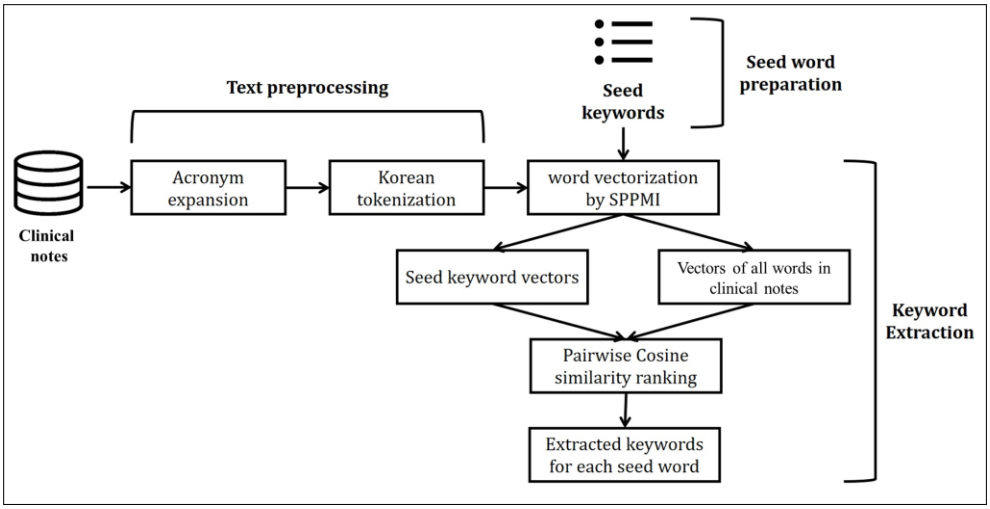

Figure 1. Overall summary of our SPPMI-based keyword expansion method

SPPMI shifted positive pointwise mutual information

Figure 1은 전체 실험의 개요를 보여준다. 실험은 text preprocessing, seed word preparation, keyword extraction의 총 3단계로 진행하였다.

Text preprocessing 단계에서, 170개의 흔히 사용되는 의학 약어들을 발견하였고, 원어로 변경하였다. 영어와는 달리, 한국어는 모든 단어를 공백으로 구별하지 않는다. 한국어는 교착어이므로, Python의 soynlp[④] 패키지를 활용하여 어절 단위의 토큰화 작업을 수행하였다. Soynlp는 미리 정의된 dictionary에 의존하지 않고 한국어 단어 사이의 경계가 발생할 확률을 문자 수준에서 추정하여 경계를 찾는다.

Text preprocessing 후, 흡연 상태 관련 seed keyword의 목록을 준비하였다. Seed keyword는 다른 흡연 상태 관련 키워드를 찾기 위한 기초 역할을 하였다. 데이터셋에서 다빈도로 관찰되는 단어를 분석하여, 두 명의 전문의가 흡연 상태 관련 키워드 총 50개를 정의하였다. 생성된 seed keyword는 비흡연자 11개, 과거 흡연자 16개, 현재

④ https://github.com/lovit/soynlp

흡연자 23개였다. seed keyword는 실험의 정확성과 계산의 효율성을 위하여 unigram, bigram으로 제한하였다. 한국어 키워드의 경우, unigram과 bigram을 soynlp에서 식별된 단어로 정의하였다. Table 2는 흡연 상태에 따라 정의된 seed keyword의 예시이다.

Table 2. Examples of seed keywords.

Never smoker	Past smoker	Current smoker
s negative Never 흡연 negative smoking negative smk never 비흡연 원래 하지 smk negative negative 하지 않음 never smoker	ex quit smk ex 금연 ya quit 금연 유지 금연 중 s ex ex smoker ex quit yo yr negative quit quit p quit ya quit 음주 smoker ya	흡연 positive ppd s positive 담배 갑 smoking current smoker smk yr 흡연 smk smk positive smoker smk ppd 흡연 p 개비 한갑 반갑 하루 갑 positive ppd 갑 넌 피운다 smoking positive current

Keyword extraction 단계에서는 seed keyword와 의미상 유사한 단어를 식별하였다. 의미상 유사도 (semantic similarity)를 계산하기 위하여, SPPMI를 이용하여 seed keyword와 임상 기록 내 모든 단어를 numerical vector로 사용하였다. 이후 각 seed keyword vector와 다른 단어 vector 간의 pairwise cosine similarity를 계산하여 순위를

매겼다. Seed keyword와 가장 높은 cosine similarity를 가진 단어를 추출할 키워드로 식별하였다. SPPMI에서 각 단어는 초기에 pointwise mutual information (PMI) score의 vector로 표시된다[39]. PMI는 아래와 같이 정의한다.

$$\mathrm{PMI}(\mathrm{word}_1, \mathrm{word}_2) = \ln\left(\frac{p(word_1, word_2)}{p(word_1) \times p(word_2)}\right)$$

PMI는 함께 발생할 확률과 개별 확률을 비교하여 두 단어 간의 연관성에 대한 확률론적 측정을 제공한다[39]. 단어 유사성을 효과적으로 계산할 수 있기 때문에, PMI 또는 normalized PMI는 자연어처리에 자주 사용된다.

SPPMI는 아래와 같이 정의한다.

$$\mathrm{SPPMI}(\mathrm{word}_1, \mathrm{word}_2) = \max\left(PMI(word_1, word_2) - \log k, 0\right)$$

SPPMI는 global constant (log k)에 따라 각 단어의 PMI 값들을 이동(shift)시킨다[40]. 저차원 단어 벡터의 경우, singular value decomposition (SVD)와 같은 행렬 분해 (matrix factorization)가 추가적으로 적용된다. k값 또는 SVD에 사용된 단일 값의 수에 따라서 SPPMI는 다른 neural network 기반의 word representation method 보다 단어 유사도를 더 잘 찾아낸다.

기존 연구에 따르면 Word2Vec과 SPPMI의 기본 메커니즘은

동일하며[41] SPPMI는 실험적으로 Word2Vec과 유사한 수준의 성능을 보인다. 본 연구에서 SPPMI를 제안하는 이유는 기존 자연어처리 연구에서 많이 사용되었던 Word2Vec과 이론적으로 동일한 메커니즘을 가지며, Word2Vec을 직접 학습하기에는 데이터가 적을 때 비슷한 방식으로 학습을 가능하게 해준다는 장점이 있기 때문이다[40,42]. 데이터셋이 작은 경우, 두 단어가 동시 발생할 확률이 아주 드물거나, 아예 등장하지 않는 단어인 경우 -∞로 계산될 수 있어 양의 점별 상호정보량 (positive PMI)으로 변환한 후, Word2Vec 공간을 최대한 근사하기 위해 parameter 값으로 log k를 사용하여 PMI 공간을 이동시켜 성능을 향상시킨다. 본 연구에서는, 동일한 값의 k (1,5,15)를 두고, singular value의 수는 (100,500,1000) 세 가지로 두고 실험을 하였다.

4) 실험 설계

SPPMI 기반 키워드 추출의 성능을 객관적으로 측정하기 위하여, 본 연구에서는 word co-occurrence, PMI vector, NPMI vector, PMI score, NPMI score, Word2Vec 총 6개의 서로 다른 모델들과 비교하였다. Text preprocessing, seed word preparation 단계는 동일하게 진행하고, keyword extraction 단계에서 각 6개의 모델을 사용하였다.

Word co-occurrence, PMI vector, NPMI vector 모델은 각 단어를 dataset의 다른 모든 단어와 함께 각각의 동시 발생 횟수 (co-occurrence counts), PMI score 및 NPMI score의 vector로 나타낸다. 이 모델들은 vector 표현을 기반으로, baseline model은 seed

keyword와 동일한 흡연 상태에 속하는 키워드를 추출하기 위해 pairwise cosine similarity 순위를 매긴다.

PMI score, NPMI score 기반 키워드 추출 모델의 경우 word vector를 생성하지 않는 대신에, 각 seed keyword와 다른 모든 단어 사이의 pairwise PMI와 NPMI score를 계산하였다. 점수의 순위를 매겨, PMI와 NPMI score 모델은 유사하게 각 seed keyword에 대한 관련 키워드를 추출한다.

Word2Vec model은 각 word vector는 주변 단어가 주어진 단어를 예측하도록 훈련된 neural network의 가중치로 표현된다. Word2Vec은 가장 기본적인 단어 임베딩 방법 중의 하나로, 다양한 단어 수준의 NLP 작업을 수행하는데 널리 사용되고 있다[43]. 본 실험에서는 서로 다른 hyperparameter의 9개 모델을 실험하였다. Word vector의 dimension으로는 100, 200, 300을 두고, context size는 2,4,6으로 지정하였다. Word2Vec으로 관련 키워드를 추출하기 위해 pairwise cosine similarity를 사용하였다.

제 2 절 결과

1) Keyword Extraction Precision

앞서 기술한 unsupervised keyword extraction model을 사용하여 총 50개의 seed keyword와 비교를 진행한 후 문서 내에서 유사도가 높은 상위 1, 5, 10, 20개의 keyword를 추출하였다. 데이터셋 내의 흡연 관련 키워드에 대한 정확한 정보는 알 수가 없으므로, 정밀도(precision)를 사용하여 키워드 추출의 성능을 측정하였다. Precision을

비교하기 위해 2명의 전문의가 독립적으로 추출한 키워드가 seed keyword와 동일한 흡연상태를 설명하는지를 평가하였다. 두 전문의 모두가 동일한 흡연 상태를 설명한다고 동의한 경우에만, 키워드 추출 결과가 정확한 것으로 간주하였다. Table 3을 보면, SPPMI 기반 keyword expansion 방법이 가장 우수한 precision을 보여줌을 알 수 있다. 문서 내 흡연과 관련된 키워드가 한정되어 있기 때문에 추출한 키워드의 수가 증가함에 따라 precision은 감소하는 양상을 보였다.

Table 3. Comparison of precision between SPPMI-based keyword expansion and six baseline models (word co-occurrence, PMI vector, NPMI vector, PMI score, NPMI score, Word2Vec).

Number of keywords / Methods	Top 1	Top 5	Top 10	Top 20
Word co-occurrence	38.00%	35.60%	30.40%	29.50%
PMI vector	42.00%	36.40%	31.20%	27.40%
NPMI vector	40.00%	35.60%	32.00%	27.70%
PMI score	38.00%	36.80%	37.00%	**34.50%**
NPMI score	36.00%	38.40%	37.40%	34.20%
SPPMI ($k = 1$)	42.00%	36.40%	31.20%	27.40%
SPPMI ($k = 5$)	38.00%	34.80%	31.80%	27.40%
SPPMI ($k = 15$)	36.00%	30.40%	27.00%	27.20%
SPPMI-SVD ($k = 1, d = 100$)	46.00%	40.80%	**38.40%**	31.50%
SPPMI-SVD ($k = 1, d = 500$)	46.00%	**41.20%**	35.40%	29.80%
SPPMI SVD ($k = 1, d = 1000$)	50.00%	38.80%	31.20%	29.50%
SPPMI-SVD ($k = 5, d = 100$)	**56.00%**	37.60%	32.00%	29.30%
SPPMI-SVD ($k = 5, d = 500$)	42.00%	36.00%	32.80%	30.00%
SPPMI- SVD ($k = 5, d = 1000$)	42.00%	34.40%	32.80%	30.10%
SPPMI SVD ($k = 15, d = 100$)	**56.00%**	40.40%	33.40%	30.60%
SPPMI- SVD ($k = 15, d = 500$)	44.00%	32.40%	29.80%	29.30%

SPPMI-SVD (k = 15, d = 1000)	44.00%	33.20%	31.00%	28.50%
Word2Vec (c = 2, d = 100)	10.00%	6.80%	7.60%	7.20%
Word2Vec (c = 4, d = 100)	11.11%	8.00%	6.20%	5.20%
Word2Vec (c = 6, d = 100)	8.82%	5.20%	4.60%	4.00%
Word2Vec (c = 2, d = 200)	10.00%	9.60%	8.40%	7.20%
Word2Vec (c = 4, d = 200)	9.09%	7.20%	6.40%	4.90%
Word2Vec (c = 6, d = 200)	8.00%	4.80%	4.60%	4.20%
Word2Vec (c = 2, d = 300)	16.00%	9.60%	7.40%	5.90%
Word2Vec (c = 4, d = 300)	8.00%	6.80%	5.20%	3.80%
Word2Vec (c = 6, d = 300)	8.00%	5.20%	4.60%	4.60%

The values of d represent the number of singular values used in the SVD or the dimension of word vectors in Word2Vec. The values of c indicate the context size used in training Word2Vec.

PMI pointwise mutual information; NPMI normalized pointwise mutual information; SPPMI shifted positive pointwise mutual information; SVD singular value decomposition

Word2Vec이 낮은 정밀도를 보이는 것이 흥미로운데 한국어와 영어가 혼용된 임상기록 기반의 데이터셋에서는 단어 임베딩 기반 방법론이 잘 작동하지 않을 수 있다. 가장 중요한 이유는 데이터 크기에 비해 많은 고유단어 (unique word)를 갖고 있기 때문이다. 의사가 진료 시 전자의무기록에 노트를 기록할 때, 대부분 환자를 진료하는 동시에 작성을 하게 된다. 이러한 실시간적 특성 때문에, 대부분의 노트는 길이가 짧고, 주로 구나 절의 형태로 구성되며, 많은 약어를 포함하고 있다. 또한 기록하는 의사마다 자주 사용하는 표현이나 기록 방식이 다른 경우가 많다. 이로 인해 의미론적으로는 유사하지만 구조적으로는 다른 용어가 데이터셋 전체에 널리 퍼져 있게 된다. 이러한 용어를 정규화하는 작업을 거치지 않으면 단어 임베딩을 일반화하여 적용하기가 어렵다. 이종언어 (bilingual) 정규화는 또 다른 중요한 향후 연구

주제일 것이다. Table 4는 SPPMI를 사용하여 추출된 키워드 예시이다. 한국어와 영어 키워드가 동시에 추출된 것을 살펴볼 수 있다.

Table 4. Examples of extracted keywords.

Never Smoker	Past Smoker	Current Smoker
흡연 negative[1] s negative never smoker	년전 ppd[2] smoking ya quit since	아직 담배[3] 못 끊었어요[4] still smoking

2) Smoking status classification

SPPMI 키워드 추출 방법으로 임상기록의 흡연 상태를 분류할 수 있다. 선행 연구[44-49]를 살펴보면, 임상기록에서 흡연 상태를 분류하기 위한 방법으로 주로 머신러닝 알고리즘이나 통계 분석을 활용하였다. 이러한 연구들에서, unigram 및 bigram bag of word로 트레이닝한 linear support vector machine (SVM)은 지속적으로 가장 높은 분류 정확도를 보여주었다[31,50-52]. SPPMI에서 추출한 키워드로 unigram, bigram bag of words를 전처리함으로써 흡연상태 분류 정확도를 더욱 향상시킬 수 있다. 각 임상기록 내 모든 단어들의 빈도를 살펴보는 대신, bag of keyword를 제안하였으며 문서 내 모든 비키워드 (non-keywords)는 동일한 단어로 취급하였다. 예를 들어, SPPMI에서 추출한 5개의 키워드로 각 문서를 나타내기로 결정하면, 각 기록은 6차원(키워드용 5개, 비키워드용 1개)의 벡터가 된다.

SPPMI 기반 추출 키워드가 흡연 상태 분류에 미치는 영향을 살펴보기 위해 분류기 (classifier)를 linear SVM으로 고정하였다. 그 후, 서로 다른 임상기록을 벡터 표현으로 변환하였고, 분류의 정확도를

비교하였다. 총 4,711개의 note 중 80%는 training dataset, 나머지 20%는 test dataset으로 활용하였다. 각 임상기록에 대해 흡연 상태가 수동 주석처리 되어있으므로, classifier의 성능에 대한 다양한 벡터 표현의 영향을 비교하기 위해 macro F1-score, precision, recall을 측정하였다. Table 5-1, 5-2, 5-3 에서는 각 방법에 따른 accuracy, precision, recall을 보여준다. Accuracy, precision, recall은 모두 macro-average로 계산되었다.

Table 5-1. Comparison of overall smoking status classification accuracy. (All accuracies are reported in F1-scores)

Methods \ Number of extracted keyword	1	5	10	20
Bag of Words	88.55%	88.55%	88.55%	88.55%
LSA	35.52%	38.54%	43.56%	50.64%
LDA	15.20%	15.20%	15.20%	15.20%
SPPMI (k = 1)	88.57%	88.79%	89.39%	89.15%
SPPMI (k = 5)	**89.73%**	87.39%	89.87%	90.31%
SPPMI (k = 15)	88.91%	89.45%	89.10%	89.80%
SPPMI-SVD (k = 1, d = 100)	88.78%	88.13%	88.38%	89.30%
SPPMI-SVD (k = 1, d = 500)	89.24%	89.25%	89.21%	90.38%
SPPMI SVD (k = 1, d = 1000)	88.70%	89.53%	**90.66%**	**91.49%**
SPPMI-SVD (k = 5, d = 100)	88.73%	**89.84%**	88.02%	89.50%
SPPMI-SVD (k = 5, d = 500)	88.89%	89.41%	89.62%	90.25%
SPPMI- SVD (k = 5, d = 1000)	88.75%	89.11%	90.47%	89.80%
SPPMI SVD (k = 15, d = 100)	89.40%	88.53%	88.12%	89.85%
SPPMI- SVD (k = 15, d = 500)	88.35%	88.65%	89.37%	89.01%
SPPMI-SVD (k = 15, d = 1000)	89.10%	88.99%	89.56%	89.31%

LSA latent semantic analysis; LDA latent Dirichlet allocation; SPPMI shifted positive pointwise mutual information; SVD singular value decomposition

Table 5-2. Comparison of overall smoking status classification precision.

Methods \ Number of extracted keyword	1	5	10	20
Bag of Words	**90.07%**	**90.07%**	90.07%	90.07%
LSA	34.65%	37.68%	41.37%	52.54%
LDA	10.92%	10.92%	10.92%	10.92%
SPPMI (k = 1)	88.37%	89.53%	89.94%	89.48%
SPPMI (k = 5)	89.87%	87.82%	90.20%	90.63%
SPPMI (k = 15)	88.61%	89.56%	89.24%	90.01%
SPPMI-SVD (k = 1, d = 100)	88.42%	88.62%	88.45%	89.48%
SPPMI-SVD (k = 1, d = 500)	88.86%	89.39%	89.46%	90.65%
SPPMI SVD (k = 1, d = 1000)	88.51%	89.57%	90.74%	**91.52%**
SPPMI-SVD (k = 5, d = 100)	88.37%	90.00%	88.32%	90.01%
SPPMI-SVD (k = 5, d = 500)	88.53%	89.53%	89.75%	90.67%
SPPMI- SVD (k = 5, d = 1000)	88.53%	89.17%	**90.87%**	90.07%
SPPMI SVD (k = 15, d = 100)	89.53%	88.69%	88.59%	90.47%
SPPMI- SVD (k = 15, d = 500)	87.96%	88.92%	89.58%	89.40%
SPPMI-SVD (k = 15, d = 1000)	88.82%	89.83%	89.79%	89.54%

LSA latent semantic analysis; LDA latent Dirichlet allocation; SPPMI shifted positive pointwise mutual information; SVD singular value decomposition

Table 5-3. Comparison of overall smoking status classification recall.

Methods \ Number of extracted keyword	1	5	10	20
Bag of Words	87.46%	87.46%	87.46%	87.46%
LSA	38.62%	41.41%	46.12%	51.01%
LDA	25.00%	25.00%	25.00%	25.00%
SPPMI (k = 1)	88.88%	88.47%	89.14%	88.97%
SPPMI (k = 5)	89.72%	87.26%	89.83%	90.23%
SPPMI (k = 15)	89.35%	89.54%	89.21%	89.82%
SPPMI-SVD (k = 1, d = 100)	89.32%	87.99%	88.53%	89.36%
SPPMI-SVD (k = 1, d = 500)	**89.79%**	89.35%	89.22%	90.26%
SPPMI SVD (k = 1, d = 1000)	89.03%	89.66%	**90.73%**	**91.59%**
SPPMI-SVD (k = 5, d = 100)	89.22%	**89.91%**	87.93%	89.28%
SPPMI-SVD (k = 5, d = 500)	89.43%	89.49%	89.71%	90.13%
SPPMI- SVD (k = 5, d = 1000)	89.10%	89.22%	90.38%	89.76%
SPPMI SVD (k = 15, d = 100)	89.43%	88.51%	87.99%	89.53%
SPPMI- SVD (k = 15, d = 500)	88.94%	88.58%	89.39%	88.95%
SPPMI-SVD (k = 15, d = 1000)	89.55%	88.52%	89.54%	89.31%

LSA latent semantic analysis; LDA latent Dirichlet allocation; SPPMI shifted positive pointwise mutual information; SVD singular value decomposition

기존 연구에서 사용되었던 unigram 및 bigram 기반의 bag of words 방법과 비교해 SPPMI 방법을 활용하면 accuracy (F1-score)가 최대 2.94% 향상되는 것을 알 수 있다 (Table 5-1). Precision과 recall 역시 SPPMI 방법을 사용하였을 때 각각 최대 91.52%, 최대 91.59%로 가장 높은 결과를 보여준다 (Table 5-2, Table 5-3). 또한

accuracy의 향상은 각 흡연 상태의 분류 정확도에서 더욱 분명해진다. Never smoker 분류 시, bag of words 대비 accuracy가 1.68% 향상되었으며 (Table 6-1), past smoker 분류 시에는 0.98% (Table 6-2), current smoker 분류 시에는 9.04%가 향상 (Table 6-3) 되었다. 또한 향상된 성능의 분류 결과는, 본 연구의 방법론이 seed keyword에서 의미 있는 다른 keyword 로의 확장이 가능함을 시사한다.

본 연구에서 사용한 방법은 latent semantic analysis (LSA)[53,54] 및 latent dirichlet allocation (LDA)[55]와 같은 기존의 다른 차원 축소 방법보다 성능이 뛰어났다. 본 연구의 우수한 흡연 분류 결과를 통해 흡연 상태와 실제로 관련된 키워드를 확장하여 추출하는 우리의 접근 방식의 높은 성능을 다시 한 번 확인할 수 있다.

Table 6-1. Comparison of never smoker classification accuracy. (All accuracies are reported in F1-scores)

Methods \ Number of extracted keyword	1	5	10	20
Bag of Words		91.57%		
SPPMI(k=1)	92.37%	91.63%	91.86%	92.37%
SPPMI(k=5)	91.33%	90.61%	91.29%	92.21%
SPPMI(k=15)	91.14%	90.61%	91.51%	91.89%
SPPMI-SVD (k=1, d = 100)	92.47%	91.10%	92.08%	90.64%
SPPMI-SVD (k=1, d = 500)	91.37%	90.87%	91.29%	92.60%
SPPMI SVD (k=1, d = 1000)	91.14%	91.40%	92.63%	**93.25%**
SPPMI-SVD (k=5, d = 100)	91.33%	91.48%	91.82%	91.63%
SPPMI-SVD (k=5, d = 500)	91.14%	91.63%	92.05%	91.82%
SPPMI-SVD (k=5, d = 1000)	91.56%	91.56%	92.50%	91.21%
SPPMI SVD (k=15, d = 100)	90.95%	91.79%	91.29%	92.80%
SPPMI-SVD (k=15, d = 500)	90.95%	91.60%	91.82%	90.83%
SPPMl-SVD (k=15, d = 1000)	91.60%	90.99%	91.40%	91.79%

SPPMI shifted positive pointwise mutual information; SVD singular value decomposition

Table 6-2. Comparison of past smoker classification accuracy. (All accuracies are reported in F1-scores)

Methods \ Number of extracted keyword	1	5	10	20
Bag of Words		89.74%		
SPPMI(k=1)	87.92%	87.66%	88.72%	88.66%
SPPMI(k=5)	88.72%	86.01%	89.46%	90.13%
SPPMI(k=15)	88.66%	88.43%	87.14%	88.48%
SPPMI-SVD (k=1, d = 100)	88.14%	86.67%	86.53%	89.12%
SPPMI-SVD (k=1, d = 500)	89.06%	87.96%	88.14%	90.21%
SPPMI SVD (k=1, d = 1000)	88.02%	89.46%	90.54%	**90.72%**
SPPMI-SVD (k=5, d = 100)	88.60%	89.06%	86.38%	89.29%
SPPMI-SVD (k=5, d = 500)	88.48%	89.00%	88.49%	90.18%
SPPMI-SVD (k=5, d = 1000)	88.37%	88.61%	90.08%	90.31%
SPPMI SVD (k=15, d = 100)	88.89%	87.34%	86.45%	88.44%
SPPMI-SVD (k=15, d = 500)	87.66%	87.40%	88.83%	88.89%
SPPMl-SVD (k=15, d = 1000)	88.54%	88.49%	88.49%	89.17%

SPPMI shifted positive pointwise mutual information; SVD singular value decomposition

Table 6-3. Comparison of current smoker classification accuracy. (All accuracies are reported in F1-scores)

Methods \ Number of extracted keyword	1	5	10	20
Bag of Words		77.88%		
SPPMI(k=1)	80.18%	81.19%	82.52%	80.95%
SPPMI(k=5)	84.91%	78.64%	84.21%	84.62%
SPPMI(k=15)	82.19%	83.96%	83.18%	84.51%
SPPMI-SVD (k=1, d = 100)	80.37%	80.58%	80.75%	83.02%
SPPMI-SVD (k=1, d = 500)	82.88%	84.11%	82.86%	84.36%
SPPMI SVD (k=1, d = 1000)	82.19%	83.57%	84.91%	**86.92%**
SPPMI-SVD (k=5, d = 100)	81.45%	84.51%	79.43%	82.52%
SPPMI-SVD (k=5, d = 500)	82.51%	82.46%	83.41%	84.47%
SPPMI-SVD (k=5, d = 1000)	81.65%	82.46%	85.02%	83.25%
SPPMI SVD (k=15, d = 100)	83.57%	81.13%	80.58%	83.33%
SPPMI-SVD (k=15, d = 500)	81.25%	81.90%	82.30%	81.55%
SPPMl-SVD (k=15, d = 1000)	82.73%	82.76%	83.81%	81.73%

SPPMI shifted positive pointwise mutual information; SVD singular value decomposition

3) Frequency distribution of the expanded keywords

본 연구에서 사용한 방법은 오로지 벡터 유사성에만 의존하기 때문에 자주 등장하지 않는 키워드도 추출이 가능하다. Figure 2를 보면 추출된 키워드의 약 60%가 전체 데이터에서 5회 미만으로 발생한다. 통계적 측정법[56]이나 feature importance를 활용하는 classifier[57]의 경우 드물게 발생하는 키워드를 추출하는데 효과적이지 않을 수 있다. 데이터의 양이 제한된 경우나 표준화나 일반화가 어려운 임상기록의 경우 드물게 발생하는 키워드를 추출하여 문서 분류에 활용하는 것이 핵심이며, 흡연 상태를 표현하는 다양한 방법을 포착하는데 효과적일 수 있다.

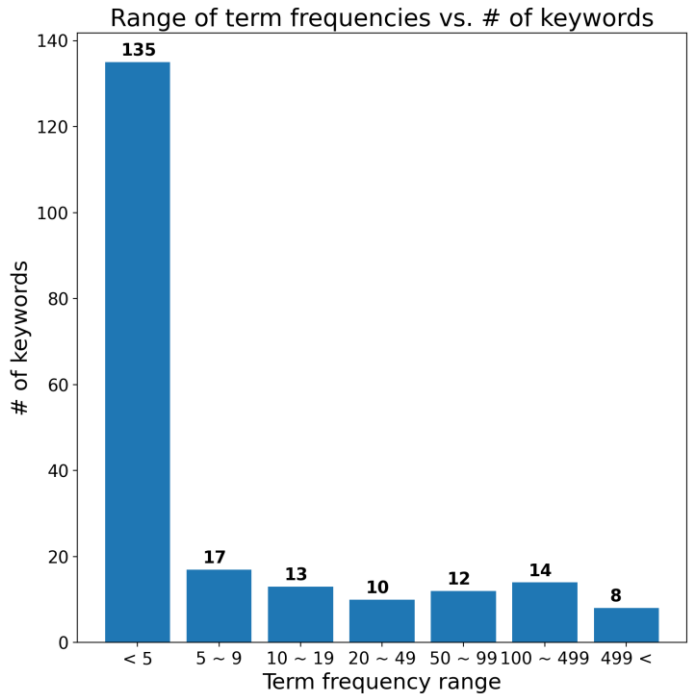

Figure 2. A number of extracted keywords with respect to their frequencies

제 3 장 Transformer를 이용한 음주 키워드 추출 및 상태분류

본 장에서는 구조화되어 있지 않은 임상기록에서 음주 관련 정보를 추출하기 위해 transformer 기반 방법론을 사용하고자 한다. Transformer 기반 언어 모델은 Wikipedia, BooksCorpus[58], CommonCrawl[59]과 같은 대규모의 텍스트 데이터셋으로 선행학습(pre-train)된다. 그 결과, 주석이 지정되어 있는 데이터셋으로 특정한 다운스트림 (downstream) 태스크 (task)를 통해 일반적인 언어패턴을 활용할 수 있으며[60], 질문 및 답변[61], 문서 분류[62], 기계 번역[63]과 같은 다양한 작업을 수행할 수 있다.

의료 분야에서 transformer 기반 언어 모델은 병원 재입원 예측[64-66], 질병 예측[67], 개념 정규화[68-70]와 같은 다양한 다운스트림 응용 모델에 활용되고 있다. 그러나 비영어권 국가 또는 2개 이상의 언어로 구성된 임상기록에서 이러한 언어 모델을 작업한 경우는 소수에 불과하다. 또한 임상기록에서 환자의 음주 관련 정보를 식별하는 기능에 대한 선행 연구는 없었다.

본 장에서는, 구조화되지 않은, 한국어와 영어가 혼용된 임상기록에서 transformer 기반 방법론으로 음주 관련 정보를 식별하여 추출하고, 음주 상태를 분류하는 연구를 하고자 한다.

제 1 절 연구 방법

1) 데이터 확보

2017년 1월 1일~ 2018년 12월 31일, 2021년 1월 1일~ 2021년 12월 31일까지 서울대학교병원 가정의학과 외래에 당뇨병 상병으로 내원한 환자의 외래초진 또는 외래경과 기록을 추출하여 사용하였다. 데이터 추출에는 2장과 동일하게 서울대학교병원 clinical data warehouse 인 SUPREME을 사용하였다. Table 7은 연도별로 연구에 활용한 임상기록의 수와 환자의 수를 나타낸 것이다. 한 환자가 여러 번 외래 진료에 방문하게 되므로, 최종 연구에는 1,982명의 4,996개의 임상기록을 활용하였다.

Table 7. Summary of basic patient statistics.

Year	Number of unique patients	Number of clinical notes	Average patient age	Number of males	Number of females
2017	1,122	2,000	64.48	619	503
2018	1,051	1,996	64.85	577	474
2021	589	1,000	66.46	319	270
Total	1,982	4,996	65.01	1068	914

시간 의존적인 연관성을 제거하기 위해, 2017, 2018년의 데이터셋은 training set으로, 2021년은 validation set으로 활용하였다.

2) 음주 상태의 정의 및 주석 (annotation) 작업

음주 상태 분류에 transformer 기반 방법을 사용하기 위해서는 pre-train을 위한 음주 관련 정보 주석 처리가 선행되어야 한다. 이에 환자의 현재 음주 상태 (current drinker, past drinker, never drinker), 음주 정도 (light drinker, heavy drinker), 폭음 유무 (binge

drinking)를 레이블링 하였다. 환자의 음주 상태에 따라, 각 임상기록을 '현재 음주자 (current drinker)', '과거 음주자 (past drinker')), '비음주자 (never drinker)', '정보 없음 (absence of information')의 네 카테고리로 분류하였다. 현재 음주자의 경우, 적정 음주자 (light drinker, 남성의 경우 매일 4잔 미만, 여성의 경우 매일 2잔 미만 음주하는 경우), 과음주자 (heavy drinker, 남성의 경우 매일 4잔 이상, 여성의 경우 매일 2잔 이상 음주하는 경우) 두 가지로 분류하였다. 과음주자 (heavy drinker)와 폭음주자 (binge drinker)의 기준은 National Institute on Alcohol Abuse and Alcoholism (NIAAA)[71]의 가이드라인을 참조하였다.

임상 경력이 있는 5명의 연구원이 annotator의 역할을 하였다. 음주 상태에 대해 수기 주석 작업을 진행하였으며, 주석 작업 첫 단계에서는 100개의 동일한 데이터셋을 레이블링 하게 하여, 일치도를 파악하였다. 레이블링 일치도는 Fleiss' kappa(κ)[72]로 측정하였으며, κ =0.9285로 높은 일치도를 보였다. 수동 주석 작업 시, 음주 상태를 분류할 뿐 아니라 각 임상기록에서 음주와 관련된 문장이나 단어, 어구도 함께 추출하여 fine-tuning에 활용할 수 있도록 하였다.

3) 텍스트 전처리

두 명의 전문의가 임상기록 내의 170개의 약어를 식별하고, 임상기록 내 자주 쓰는 기호와 두문자어를 단어 또는 완전한 표현으로 대체하였다. 간단한 문자열 교체만으로 전처리가 가능하기 때문에 대규모 외부 데이터 리소스를 사용하는 것 보다 계산 속도가 더 빠르다. 몇 가지 기호 및 두문자어의 예시를 figure 3에서 확인할 수 있다.

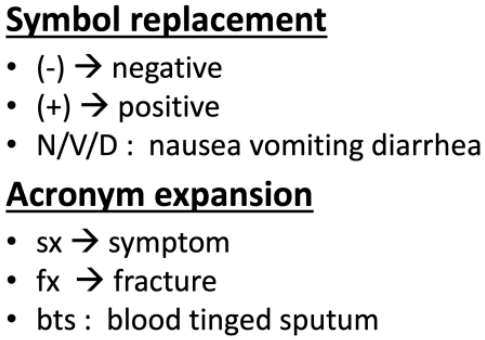

Figure 3. A few examples symbols and acronyms preprocessed during the symbol replacement and acronym expansion.

4) 음주 정보 추출 실험 방법

실험은 크게 두 단계로 나누어 진행하였다. 첫 번째 단계에서는, 임상기록 내 음주 관련 문장을 추출하여, 음주 관련 정보가 있는 문서를 찾아내었다. 두 번째 단계로는 음주 관련 문장 또는 키워드가 있는 문서를 대상으로, 음주 행태에 대한 상세 분류 (light drinker, heavy drinker, past drinker, binge drinker)를 진행하였다.

임상기록 내 음주 관련 정보 추출에는 XLM-RoBERTa (cross-lingual language model - robustly optimized BERT pretraining approach)[73]를 사용하였다. Figure 4는 음주 정보 추출을 위한 방법론에 대한 모식도이다. XLM-RoBERTa는 대규모 transformer 기반 언어 모델[74]로서, XLM-RoBERTa는 문맥화된 단어 임베딩을 학습하기 위해 RoBERTa와 동일한 transformer 구조를 사용한다. RoBERTa는 Byte-Pair Encoding[75,76]을 사용하여 입력 문장 내의 단어를 하위 단어로 토큰화하여 모델이 어휘 외 단어와 희귀 단어를 보다 효과적으로 처리할 수 있도록 한다. 그 후 입력 문장의 하위 단어는 12개의 양방향

(bidirectional) transformer 레이어를 통해 처리되며, 각 레이어는 attention mechanism을 적용한다. Attention mechanism을 기반으로 입력 문장의 다른 하위 단어에 대한 각 단어의 영향이 다르게 평가되어, 각 하위 단어를 encoding하고 삽입하여 입력 문장 내에서 문맥적으로 의미있는 다른 하위 단어를 통합할 수 있다. XLM-RoBERTa는 RoBERTa와 동일한 모델 구조를 가지고 있으며, pre-train 시 multilingual model objective[73,77]를 사용한다. 단일언어 데이터셋의 input으로부터 마스킹된 단어를 예측하는 대신, XLM-RoBERTa는 다국어 데이터셋으로부터 input 문장을 받아 마스킹된 단어를 예측한다.

XLM-RoBERTa는 Wikipedia corpus의 104개 언어로 사전 학습되며, 한국어와 영어 모두를 지원하여 한영 혼용 이종언어 임상기록에 대한 실험에 적절하다. RoBERTa에 비해 다중 언어로 구성된 임상기록을 fine-tuning하는데 더 강점이 있다. 본 연구의 데이터셋의 임상기록은 한국어와 영어만 포함되어 있기 때문에 본 실험은 이종언어 환경에서 XLM-RoBERTa의 추출 성능을 검증하는 것이 그 목적이다. XLM-RoBERTa는 임상기록의 음주 관련 정보와 다른 정보를 구별하는 분류 기능이 없기 때문에, 출력 레이어 (output layer) 위에 768개의 hidden node가 있는 추가적인 선형 레이어를 쌓는다. 본 연구에서는 XLM-RoBERTa의 output layer에 사전 기술한 음주 정보를 annotation한 데이터셋을 사용하였으며, XLM-RoBERTa와 분류 레이어를 5 epoch의 학습과 early stopping을 진행했다. 학습 속도 (learning rate)가 1 X 10^{-5}이고 가중치 감쇄 (weight decay)가 0.01인 AdamW optimizer[78]를 training에 사용하였다. 이에 training 동안, linear layer optimization 뿐 아니라 cross-entropy loss function까지 최소화하기

위해 사전학습된 transformer layer도 업데이트된다. 본 연구의 모든 실험은 HuggingFace [⑤] 에서 다운로드한 사전훈련된 XLM-RoBERTa 가중치와 함께 PyTorch[⑥]를 사용하였다.

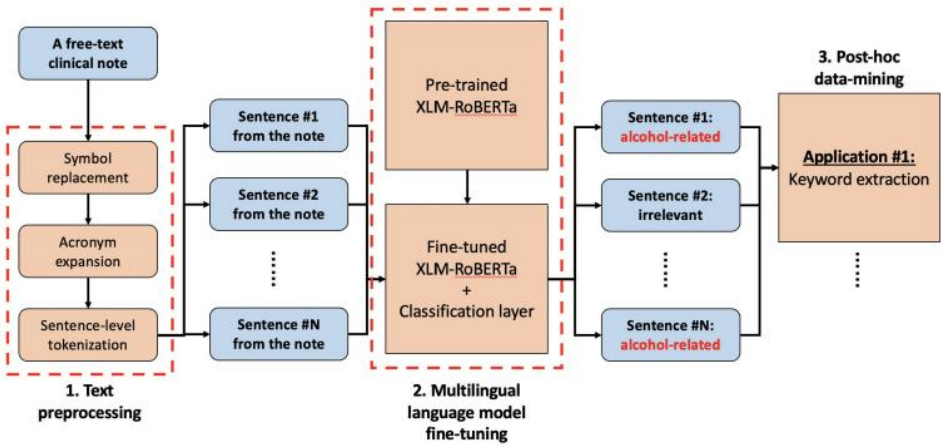

Figure 4. An overview of the transformer-based language model for extracting patients' alcohol-related information from their free-text clinical notes.

XLM-RoBERTa cross-lingual language model - robustly optimized BERT pretraining approach

5) 음주 상태 분류 실험 방법

기존 연구들은 대부분 임상기록 중 음주 상태를 자동 분류하기 위해 Bag of Words (BoW) 기반 모델을 채택하여 임상기록 내 모든 단어를 사용하는 경우가 대부분이었다[34-37]. 반면, 본 연구에서는 대규모 attention 기반 언어 모델을 채택하여 모델 성능을 향상시켰고, 기록 내 중요한 음주 관련 정보를 자연어처리 모델에 직접 주입하는 방법을 사용하였다 (Figure 5). 본 모델에서 환자의 임상기록은 초기에 사전

⑤ https://huggingface.co/xlm-roberta-large
⑥ https://pytorch.org/

훈련된 attention 기반 언어 모델의 input으로 사용된다. 전체 임상기록에서 의미론적 정보를 수집하면서 CLS token (special classification token) vector에 특이적인 환자 수준의 음주 관련 정보를 확장한다. 환자 수준의 음주 정보로는, 주종과 음주 소비량이라는 두 가지 정보를 추가로 포함하였다. 데이터셋에서 환자가 소비하는 주종을 추출하기 위해 환자의 음주 소비 패턴 정보가 담긴 임상기록 내에서 자주 발생하는 단어를 나열했다. 전체 데이터셋을 리뷰하여 환자들이 자주 마시는 주종을 12가지로 좁혔다 (Figure 5). 12가지 주종의 목록에 해당하는 문자열 매칭을 적용하면 음주 패턴 예측이 가능하다. 이 문자열 매칭 결과를 바탕으로 one-hot encoding을 사용하여 음주 소비 유형을 나타낼 수 있다. 음주 소비량은 수동 annotation을 통해 정보를 추출했다. 5명의 annotator가 각 임상기록의 음주량과 폭음 상태를 레이블링 할 때 일일이 소비량에 대한 기록 내 단어나 어구도 추가로 레이블링 하였다. CLS token이 환자의 보조적인 음주 관련 정보 (additional information)와 연결되며 추가 정규화 및 비선형 레이어를 통해 음주 상태 분류를 진행하였다.

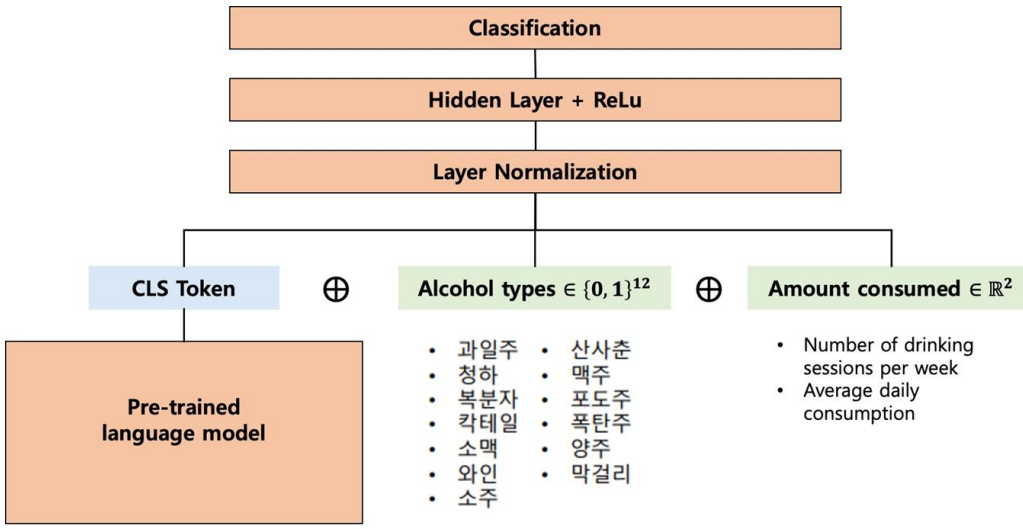

Figure 5. A summary of classification model; Multilingual + additional information

CLS token special classification token; ReLu rectified linear unit

제 2 절 결과

1) 음주 정보 추출

Table 8은 주석 작업을 완료한 dataset의 문장 개수를 정리한 것이다. 총 문장수는 139,397개였으며, 한국어와 영어가 혼용되어 있었고, 그 중 음주 관련 문장은 총 3,299개였다.

Table 8. Overview of the annotated dataset

Year	Total number of sentences	Number of English sentences	Number of Korean sentences	Number of bilingual sentences	Number of alcohol-related sentences
2017	49,044	10,684	11,857	26,503	1,318
2018	55,139	12,066	16,482	26,591	1,435
2021	35,214	7,153	11,587	16,474	546
Total	139,397	29,903	39,926	69,568	3,299

앞 절에서 제안한 XLM-RoBERTa 기반 음주 정보 추출 방법의 성능 검증을 위해 다른 자연어처리 알고리즘과 성능을 비교하였다. 임상기록에서 음주 관련 정보 추출에 대한 선행 연구[34-37,79]에서 다룬 바와 같이, 본 연구에서도 BoW를 기준으로 사용하였다. 중요도와 상관없이 발생하는 다빈도 단어를 설명하기 위해 TD-IDF를 사용하여 단어 빈도를 조정하였다. BoW를 기반으로 SVM[80], XGBoost[81] 와 Random Forest[82] 세 가지 유형의 분류기를 테스트하였으며 모두 10-fold cross-validation의 hyperparameter를 사용하였다. 또한 transformer 기반의 multilingual BERT와 multilingual distilled version of BERT (DistilBERT) 도 fine-tuning 하였다.

Table 9에서 각 알고리즘 별 추출 성능을 살펴볼 수 있다. Accuracy, precision, recall은 모두 macro-average로 계산되었다. 모든 transformer 기반 모델은 임상기록에서 음주 관련 문장을 추출하는데 좋은 성능을 보인다. Gururangan 등[60]이 지적한 바와 같이 transformer 기반 언어 모델은 fine-tuning을 통해 대규모 pre-train 과정을 기반으로 대규모 데이터셋에 빠르게 적용할 수 있다. 한편, BoW + Random Forest의 경우 모든 문장을 환자의 음주 소비와 관련이 없는 것으로 분류하여 모델의 레이블 불균형 문제를 극복하지 못했다. BoW + SVM 모델은 적절한 추출 성능을 제공하나, BoW 기반 모델은 차원이 높고 out-of-vocabulary (OOV) 문제가 있다. 문장과 단어의 수가 증가함에 따라 문장 벡터의 차원도 증가하여, 메모리 이슈가 발생할 수 있다. 더욱이 학습한 적이 없는 새로운 단어가 추가되면 새로운 단어를 통합하기 위한 재훈련이 필요하다. 즉, 의료기관 내에서 생성되는 대규모 임상기록을 분석할 때 BoW 기반 모델을 사용하는 것은 상당한 계산 문제를 야기할 수 있다. 반면 transformer 기반 모델은 고정된 차원을 사용하여 단어를 나타내므로, 새로운 단어가 도입될 때 벡터의 차원이 확장되는 것을 방지할 수 있다. 또한 byte pair encoding[75]을 사용하기 때문에 새로운 단어를 이전에 관찰되었거나 사전 훈련된 하위 단어로 토큰화하고 처리하여 OOV 문제를 해결할 수 있다.

Table 9. Overall comparison of the extraction performance of different NLP models.

Model	Macro F1-score	Precision	Recall
XLM-RoBERTa	**84.70%**	**78.23%**	**95.33%**
Multilingual BERT	84.20%	77.61%	95.29%
Multilingual DistilBERT	83.00%	76.25%	94.86%
BoW + SVM	83.37%	77.24%	93.38%
BoW + XGBoost	60.60%	90.39%	56.30%
BoW + Random Forest	49.61%	49.23%	50.00%

XLM-RoBERTa cross-lingual language model - robustly optimized BERT pretraining approach; DistillBERT distilled version of BERT; BoW bag of words; SVM support vector machine; XGBoost extreme gradient boosting

Table 10에서는 한국어 또는 영어 혹은 한영이 혼용된 모든 경우 각 모델에 따라 비슷한 수준의 성능을 보임을 알 수 있으며, table 9의 결과와 같은 양상을 보인다. 모든 한국어 또는 영어, 혹은 한영 혼용 임상기록에서 XLM-RoBERTa 기반 추출 방법은 Multilingual BERT 및 Multilingual DistilBERT를 포함한 다른 모든 방법보다 우수한 성능을 보인다. 이는 Conneau등[73]이 발표한 결과와 일치한다. 흥미로운 점은, 한국어 문장 추출 성능이 영어나, 한영 혼용 문장 추출 성능보다 낮다는 것이다. 추출 성능이 떨어지는 이유는 토큰화 문제 때문일 수 있다. 영어와 달리 한국어는 단어 사이의 경계가 공백 (space)으로 정의되지 않는 교착어이다. 따라서 사전에 전처리 없이 공백을 기반으로 한국어 문장을 토큰화 하는 것은 XLM-RoBERTa의 추출 성능이 저하되는 원인 중 하나일 수 있다.

Table 10. The extraction performance with respect to different languages used in the dataset (measured in terms of macro F1-score)

Model	Macro F1-score (English)	Macro F1-score (Korean)	Macro F1-score (Bilingual)
XLM-RoBERTa	**94.25%**	**67.69%**	**91.01%**
Multilingual BERT	92.86%	67.06%	90.74%
Multilingual DistilBERT	93.43%	65.03%	90.02%
BoW + SVM	93.62%	66.20%	89.37%
BoW + XGBoost	49.41%	57.66%	66.58%
BoW + Random Forest	49.42%	49.81%	49.54%

XLM-RoBERTa cross-lingual language model - robustly optimized BERT pretraining approach; DistillBERT distilled version of BERT; BoW bag of words; SVM support vector machine; XGBoost extreme gradient boosting

2) 음주 관련 key concept 생성

본 연구에서 제안하는 방법은, 이종언어 임상기록에서 음주 관련 문장을 자동으로 추출하는 것 외에도 다양한 사후 분석에 활용될 수 있다. 잠재적인 사후 분석의 예로, 이종언어 임상기록에서 음주 관련 key concept을 생성해 볼 수 있다. Key concept을 생성하기 위해 XLM-RoBERTa 위의 추가 선형 계층의 출력 노드에 softmax operation을 적용하여, 각 입력 문장에 대한 예측의 모델의 신뢰 수준을 계산할 수 있다. 이를 통해 90% 이상의 신뢰도로 올바르게 예측된 음주 관련 문장을 예측할 수 있다. 이러한 음주 관련 가능성이 높은 문장 내의 모든 unigram 및 bigram에 대해 lift를 계산하였다. Lift는 아래와 같이 정의된다.

$$\text{Lift(Word w)} = \frac{\textit{Frequency of w in alcohol-related sentences}}{\textit{Frequency of w in the entire dataset}}$$

Lift 값이 클수록 특정 unigram 및 bigram이 음주 관련 문장의 특징이 더 크다는 것을 나타내므로 음주 관련 key concept으로 lift 값이 0.95보다 큰 unigram 또는 bigram을 추출하였다. Table 11에서 추출된 key concept의 몇 가지 예를 제시하였다.

Table 11. Examples of the extracted alcohol-related keywords.

Keywords	Language	Description
1000 cc	English	Refers to the volume of beer (cubic centimeter)
positive alc	English	Refers to positive “alcohol”, indicating that a patient is a drinker
alcohol social	English	Indicates that a patient is a social drinker
alc 안마심	Bilingual	Indicates that a patient doesn’t consume alcohol
생맥주	Korean	A Korean word indicating “draft beer”
한병이상	Korean	A Korean phrase indicating “more than one bottle”
포도주	Korean	A Korean word indicating “wine”

3) 음주 상태 분류

전체 데이터셋의 현재 음주 상태, 과음 유무, 폭음 유무에 따른 음주 소비 패턴에 대한 일반적 통계는 Table 12, 13, 14에 요약되어 있다. Table 12에서 알 수 있듯, 모든 연도에서 현재 음주자가 가장 많은 비율을 차지하였고, 과거 음주자가 가장 적었다.

Table 12. General statistics of current status in the dataset.

Year	Current status patterns	Number of patients	Average age	Number of males	Number of females
2017	Non-drinkers	428	65.89	157	271
	Current drinkers	499	60.38	450	49
	Past drinkers	20	64.9	20	0
2018	Non-drinkers	441	66.95	165	276
	Current drinkers	508	61.87	453	55
	Past drinkers	29	67.69	24	5
2021	Non-drinkers	232	68.25	75	157
	Current drinkers	246	64.71	220	26
	Past drinkers	12	57.83	10	2

Table 13. General statistics of amount status in the dataset

Year	Current status patterns	Number of patients	Average age	Number of males	Number of females
2017	Light drinkers	144	60.38	135	9
	Heavy drinkers	121	58.81	102	19
	No information	234	61.2	213	21
2018	Light drinkers	262	61.48	229	33
	Heavy drinkers	32	59.63	32	0
	No information	214	62.69	192	22
2021	Light drinkers	93	63.84	86	7
	Heavy drinkers	13	62.85	13	0
	No information	140	65.46	121	19

Table 14. General statistics of binge status in the dataset

Year	Current status patterns	Number of patients	Average age	Number of males	Number of females
2017	Binge drinkers	131	58.64	116	15
	Non-binge drinkers	143	60.88	128	15
	No information	225	61.08	206	19
2018	Binge drinkers	157	59.04	144	13
	Non-binge drinkers	264	63.38	223	41
	No information	87	62.4	86	1
2021	Binge drinkers	49	61.57	48	1
	Non-binge drinkers	56	65.55	50	6
	No information	141	65.46	122	19

임상기록에서 음주 소비 패턴을 자동으로 추출하기 위해 다음과 같은 단계를 거쳤다. 먼저, 각 임상 노트에 있는 모든 문장을 attention 기반 언어 모델의 input으로 사용하였다. 이후 음주 형태, 최대 음주량, 평균 음주량에 대한 정보가 포함된 임상기록의 경우, 이 보조 정보를 encoding 하여 attention 모델의 output vector에 추가하였다. 이후, 수정된 vector를 각 임상 노트의 현재 음주 상태, 음주량, 폭음 상태를 분류하는 linear layer로 활용하였다.

본 연구에서 제안하는 모델 (Figure 5)의 성능을 검증하기 위해, vanilla attention 기반 언어 모델과 성능을 비교하였다. Macro F1-score를 사용하여 성능을 측정하였으며, 최근 제안된 attention 기반 언어모델 중 다중 언어를 처리할 수 있는 multilingual BERT[77]와 XLM-RoBERTa[73]를 사용하였다. 음주 상태 분류에 대한 성능은 table 15에서, 음주 양 (amount status) 분류에 대한 성능은 table 16에서,

폭음 유무 (binge status) 분류에 대한 성능은 table 17에서 확인할 수 있다. 본 연구에서 제안하는 모델은, 세가지 음주 소비 행태 모두에서 우수한 분류 성능을 보였다. 또한 사용하는 기본 언어 모델의 유형과 상관없이, 보조 음주 정보가 포함되면, 언어 모델의 추출 성능이 향상된다. Table 15에서 현재 음주 상태에 대한 Macro F1-score를 살펴보면, multilingual BERT의 성능은 4.0%, XLM-RoBERTa의 성능은 각각 3.4% 증가하였다. Precision과 recall 모두 성능이 향상됨을 확인할 수 있다. 적정 음주자와 과음주자를 구별하는 음주 양에 대한 분류를 살펴보면 multilingual BERT + additional information 이 F1-score 80.5%, recall 82.2%, XLM-RoBERTa + additional information의 경우 precision 88.8%로 additional information을 함께 모델에 적용한 경우 성능의 향상을 보였다 (Table 16). Table 17에서 확인할 수 있듯 폭음 상태를 살펴보는 경우에서도 XLM-RoBERTa + Additional information에서 F1-score 73.0%, precision 76.7%, recall 77.0%로 가장 우수한 성능을 보였다. 즉 세가지 음주 패턴에 대해 multilingual BERT는 평균 9.87%, XLM-RoBERTa는 평균 11.7%의 성능 향상을 보였다.

Table 15. Pattern extraction performance in current status

Models	F1-score	Precision	Recall
BoW + SVM	77.3%	82.0%	74.8%
BoW + RF	62.7%	61.9%	63.5%
BoW + XGBoost	81.9%	87.9%	78.6%
Multilingual BERT	82.4%	91.3%	78.2%
XLM-RoBERTa	79.0%	89.5%	74.9%
Multilingual BERT + Add (Proposed)	**86.4%**	**92.0%**	**82.8%**
XLM-RoBERTa + Add (Proposed)	82.4%	91.3%	78.2%

BoW bag of words; SVM support vector machine; RF random foreset; XGBoost extreme gradient boosting; XLM-RoBERTa cross-lingual language model - robustly optimized BERT pretraining approach; Add, additional information

Table 16. Pattern extraction performance in amount status

Models	F1-score	Precision	Recall
BoW + SVM	67.6%	66.1%	75.3%
BoW + RF	54.1%	53.3%	55.2%
BoW + XGBoost	67.9%	66.8%	69.7%
Multilingual BERT	55.4%	54.0%	57.7%
XLM-RoBERTa	53.9%	53.8%	57.0%
Multilingual BERT + Add (Proposed)	**80.5%**	81.5%	**82.2%**
XLM-RoBERTa + Add (Proposed)	77.6%	**88.8%**	74.9%

BoW bag of words; SVM support vector machine; RF random foreset; XGBoost extreme gradient boosting; XLM-RoBERTa cross-lingual language model - robustly optimized BERT pretraining approach; Add, additional information

Table 17. Pattern extraction performance in binge status

Models	F1-score	Precision	Recall
BoW + SVM	67.9%	**78.8%**	68.3%
BoW + RF	26.7%	53.8%	34.0%
BoW + XGBoost	66.2%	76.3%	65.3%
Multilingual BERT	68.8%	69.0%	74.5%
XLM-RoBERTa	65.0%	66.0%	67.3%
Multilingual BERT + Add (Proposed)	69.3%	69.3%	72.7%
XLM-RoBERTa + Add (Proposed)	**73.0%**	76.7%	**77.0%**

BoW bag of words; SVM support vector machine; RF random foreset; XGBoost extreme

gradient boosting; XLM-RoBERTa cross-lingual language model - robustly optimized BERT pretraining approach; Add, additional information

제 4 장 종합 고찰

1. 연구의 목적

본 연구는 한국어와 영어가 혼용된, 구조화 되어있지 않은 임상기록에서 흡연, 음주와 같은 유의미한 정보를 추출하고 문서를 분류하기 위한 연구이다. 현재 의료 분야에서 다양한 자연어처리 연구가 진행되고 있으나 대부분 구조화된 영상검사나 병리검사 판독문 또는 구조화된 서식 기반이 대부분이며, 한국어와 영어가 혼용된 자유진술문 기반 연구는 거의 이루어지지 않은 실정이다. 흡연과 음주는 거의 모든 질병의 위험인자로 작용하고 있으며, 진료 현장에서 필수적으로 청취하여야 하는 임상 정보이자, 의학 연구에 반드시 활용되는 중요한 변수이다. 연구 목적으로 흡연과 음주 정보를 확보하기 위해서는, 대부분 많은 비용과 인력을 들여 설문 조사를 실시하거나, 일일이 수기로 해당 정보를 추출해야 한다. 그러나 많은 경우 흡연과 음주 정보는 진료 시 시행되는 문진을 통해 확보되고 있어 임상기록을 활용하는 경우 충분히 해당 정보를 알 수 있으나, 작성하는 의료인 특이적이며 한국어와 영어가 혼용되어 있는 구조화 되어있지 않은 임상기록에서 자동으로 추출하고 상태를 분류하는 시도는 거의 이루어지지 않았다.

본 연구에서는, 두 가지 방법으로 키워드 추출 및 문서 분류를 진행하였다. 먼저, 제 2장에서는 SPPMI 기반 방법을 사용하여 흡연 정보 추출 및 분류를 진행하였다. SPPMI 모델 학습을 위하여, seed

keyword 준비 단계가 필요하였다. 제 3장에서는, transformer 기반 방법을 사용하여 음주 정보 추출 및 분류를 진행하였으며, 2장과는 달리 데이터 전처리 시 음주 관련 단어 또는 문장을 함께 수기로 추출하여, 실험의 fine-tuning 시 활용하였다.

2. 연구의 주요 결과

1) SPPMI를 이용한 흡연 키워드 추출 및 상태 분류

2장의 연구에서는, unsupervised learning의 방법 중 하나인 SPPMI를 사용하여 임상기록 내 흡연 관련 키워드 추출 및 문서 분류를 진행하였다. SPPMI 실험을 위하여 seed keyword preparation 과정이 필요하였는데, annotator가 사전에 수기로 annotation 한 흡연 상태에 따라 다빈도로 관찰되는 단어 중 전문의의 판단 하에 각 흡연 상태와 관련 있는 keyword를 확정하였다. SPPMI 실험 시 seed keyword 와 의미론적으로 유사한 단어를 계산하여, 각 문서의 keyword를 추출하였다. 이를 기반으로 문서의 흡연 상태를 분류하였을 때, SPPMI 기반 방법론이 Word2Vec, BoW, LSA, LSD보다 우수한 성능을 보였다.

2장에서는 BERT[77]나 GPT[76] 와 같은 pre-trained language model은 사용하지 않았는데, 그 이유는, 연구에서 사용한 dataset의 특성이 BERT와 GPT모델에 적합하지 않기 때문이다. Pre-trained language model의 경우 영어 기반의 corpus를 기반으로 일반 목적의 데이터셋의 연구에 효과적인 경우가 많으며, 2장의 경우와 같이 2개 이상의 언어로 구성되어 의료 특화 적인 문서의 경우 pre-train을 하기가 어렵다. 또한 본 연구에서 활용한 임상기록은, 의사가 환자를

진료하는 동시에 생성되므로 약어가 많고, 일반적으로 잘 쓰지 않는 의료 특수 용어로 구성되며, 완전한 문장이 아닌 불완전한 구나 절의 형태로 기록되는 경우가 많다. 비록 BioBERT[83]와 같은 의료 특화적인 fine-tuned language model이 있긴 하나, 이 또한 임상기록에 특화된 모델은 아니다. 무엇보다도 현재, 한국어와 영문이 혼용된 임상기록에 대한 corpus가 부재하여, pre-training 이 필요한 언어 모델을 사용하는 데에는 한계가 있으며, 이에 unsupervised learning 이 가능한 SPPMI 방법을 채택하여 실험을 하였다.

본 장의 실험에서 키워드를 추출할 때 Word2Vec이 낮은 추출 정확도를 보였으며 (최대 precision 16%), 단어간 의미 유사도를 찾아내는데 SPPMI 보다 성능이 떨어짐을 확인할 수 있었다. 더 길고 방대한 임상기록으로 구성된 대규모 데이터가 있다면 Word2Vec의 성능을 향상시킬 수 있겠지만, 대규모 데이터셋을 구성하는데 많은 비용과 인력이 소모된다는 단점이 있다. 따라서 본 장에서 제안하는 SPPMI 기반의 방법론이 적절한 대안이 될 수 있을 것이다.

2) Transformer를 이용한 음주 키워드 추출 및 상태 분류

3장의 연구에서는 transformer 기반의 언어 모델을 사용하여 음주 관련 정보를 추출하고 문서를 분류하였다. 먼저 XLM-RoBERTa를 이용하여 문서 내 음주 관련 키워드를 추출하는 실험을 진행하였다. 그 다음 단계로는 transformer 기반 방법론에 주종과 음주 잔 수 및 횟수 정보와 같은 additional information을 더하여 추가 학습을 통해 음주 상태, 과음 유무, 폭음 유무에 대해 문서 분류를 진행하였다.

본 장에서는 데이터를 전처리하는 과정에서, 음주 상태를

annotation할 뿐 아니라, 문서 내 음주 관련 정보도 함께 수기로 추출하여, 실험 모델을 fine-tuning 하는데 사용하였다. 3장의 연구에서는 transformer 기반의 자연어처리 과정에서 흔히 사용하는 concept normalization 과정을 거치지 않았는데, 그 이유는 음주 관련 정보를 나타내는 단어가 다양하고, 문서 내 적은 빈도로 관찰되기 때문이다. 본 연구에서 제안하는 방법은, 한국어와 영어가 혼용된 다양한 정보를 담고 있는 임상기록에서 특정 정보를 추출하고 문서를 분류하는데 유용할 수 있다. 의료 데이터의 정규화 또는 표준화에 대한 다양한 시도가 이루어지고 있으나, free text로 구성된 임상기록의 표준화는 쉽지 않은 실정이다. XLM-RoBERTa 실험으로 증명된 key concept 추출은, 2개 이상의 언어가 혼용된 임상기록에서 의미있는 key concept을 추출하여 텍스트 리소스를 구축하거나, 확장하는 transformer 기반 언어 모델로 발전 가능할 것이다.

문서 내 음주 관련 키워드를 추출하는 실험 이후, 음주 관련 키워드가 포함된 문서만을 추려 음주 상태, 과음 유무, 폭음 유무에 대해 문서를 분류하였다. 문서 분류 시, 키워드 추출에 사용한 transformer기반의 방법에, 음주 특화적인 키워드인 한국어로 이루어진 12개의 주종을 표현하는 단어와, 소비하는 잔 수 또는 음주 횟수에 대한 숫자 정보를 함께 문서 분류를 위한 학습에 사용하였다. 기존의 multilingual BERT만을 사용한 경우보다 additional information을 함께 학습한 경우 성능이 더 향상됨을 알 수 있었다. 즉 transformer 기반 방법으로 2개 이상 언어로 구성된 임상기록에서 pre-train 된 대규모 언어 모델을 fine-tuning 하면 downstream NLP task에서 높은 성능을 보여준다는 결론을 얻을 수 있다. 병원 재입원 예측[65,66], 질병 예측[68] 및 개념

정규화[68-70]를 위한 자유진술문으로 구성된 임상기록을 활용한 연구는 다수 진행된 바 있으나, 언어모델을 fine-tuning 하며 다국어 임상기록에서의 성능을 살펴본 연구는 거의 없었다. 또한 기록 내 음주 관련 추가 정보를 사용하여 도메인 특화적 정보를 언어 모델의 출력과 결합하면 downstream NLP 모델의 성능을 더욱 향상시킬 수 있음을 확인하였으며, 이전의 연구와 상응하는 결과를 보였다[84-86].

3. 연구의 종합적 강점

과거 연구에서는, 흡연과 음주 정보는 주로 구조화된 설문조사나 수기로 임상기록을 리뷰하는 방법으로 확보되었다. 이는 많은 비용과 인력이 소모되는 작업이며, 본 연구는 전자건강기록으로부터 흡연과 음주 정보를 쉽게 획득할 수 있는 방법에 대해 제안한다.

본 연구는 한영 혼용으로 작성된, free text로 이루어진 임상기록에서 유의미한 정보를 추출하고 문서를 분류한 최초의 연구이다. 또한, 기존 연구가 대부분 흡연자/비흡연자 또는 음주자/비음주자의 2개의 카테고리로 문서를 분류한 한 것에 반해, 본 연구는 흡연을 현재 흡연, 과거 흡연, 비흡연 3가지 카테고리로 나누었으며, 음주의 경우 현재 음주, 과거 음주, 비음주를 비롯 NIAAA 기준에 맞는 과음주자, 폭음주자까지 분류해낼 수 있는 문서 분류 알고리즘을 제안한다. 흡연, 음주와 같은 유해한 생활습관은 그 누적 행태와 사용량 등이 매우 중요한 요소이므로, 다양한 카테고리로 문서를 분류할 수 있는 기능을 제공하는 큰 강점을 지녔다 할 수 있겠다.

본 연구에서 제안하는 SPPMI 기반 방법은 문서 분류 시 최대 accuracy (macro F1-score 기준) 91.49%, multilingual BERT +

additional information 기반 방법은 86.4%를 보여 높은 수준의 정확도를 보여주었다. 임상기록은 하나의 문서에 흡연 음주와 같은 생활습관 정보는 물론, 주 호소, 주관적 건강상태, 객관적 건강상태, 계통적 문진, 신체 진찰, 동반 질환, 알러지 유무, 현재 복용 중인 약물, 과거 병력, 가족력, 통증 유무, 추정 진단, 향후 계획 등 다양하고 방대한 정보를 포함하고 있다. 따라서 하나의 임상기록을 구성하는 총 단어수가 방대하여도 실제로 흡연 또는 음주상태를 기술하는 단어는 1-2개에 그치는 경우가 많다. 또한 흡연 관련 키워드 중 60% 이상이 전체 데이터 중 5회 미만으로 출현하기 때문에 일관적인 concept normalization을 적용하기 어렵다. 이에 pre-train을 위한 corpus가 부재한 상황에서는 unsupervised learning의 일환인 SPPMI 기반 키워드 추출을 통해 상태 별 키워드를 확장하고, 문서 분류를 진행해볼 수 있겠다. 한편 음주의 경우, 흡연 보다 임상 정보가 복잡한 경우가 많은데, 음주량, 종류, 횟수, 빈도 등 문진으로 수집하는 데이터의 종류가 흡연보다 다양하기 때문이다. 게다가 과음 및 폭음을 분류하는 기준도 성별에 따라 상이하며, 표준 잔 기준으로 알코올 양(g)을 계산해야 하는 번거로움이 있다. 이에 본 연구에서는 최근 대두되어 높은 성능을 자랑하는 transformer 기반의 방법을 음주 정보 추출 및 문서 분류에 적용하였으며, pre-train을 위한 사전에 마련된 corpus가 부재하므로, annotation 과정에서 음주 관련 정보를 함께 수기로 추출하여 모델 성능 향상에 활용하였다.

마지막으로, 본 연구에서 제안하는 방법은 임상기록 내 흡연 또는 음주 정보가 없는 문서를 선별할 수 있다는 장점이 있다. 환자의 일차예방 및 이차예방을 위해 진료현장에서 흡연과 음주력을 필수로

문진하도록 권고되고 있다. 이전 임상 기록을 토대로, 해당 정보가 계속 존재하지 않는 경우 적절한 알람을 임상의에게 제공하여, 환자에게 흡연 음주력을 문진하고, 기록하여 적절한 조치를 할 수 있도록 유도할 수 있다.

4. 연구의 한계

본 연구는 몇 가지의 한계점이 있다. 우선, 의사가 환자를 진찰할 때 흡연이나 음주력과 같은 생활습관을 문진하였으나, 기록하지 않았을 수 있다. 이러한 경우, 실제로 정보는 획득되었으나 임상기록 내에는 해당 정보가 부재하므로, 본 연구에서 분류한 '정보 없음'으로 분류될 가능성이 있다. 또한, 본 연구는 서울대학교병원의 가정의학과와 내과 일부의 임상기록만 활용하였기 때문에, 다른 진료과나 의료기관으로의 확장이 어려울 수 있다. 향후 다양한 기관의 임상기록을 활용한 타당도 검증이 필요할 것이다.

한편, 수동 주석 과정에서의 오류나, SPPMI 사용시의 seed keyword 선정과 같은 수작업으로 이루어지는 전처리 과정에서의 오류가 실험 성능에 영향을 줄 수 있다. 이를 위해서는 2명 이상의 숙련된 전문가가 필요하고, 같은 데이터셋으로 수동 작업을 한 후 얼마나 결과가 일치하는지 평가하는 단계가 중요할 것이다. 본 연구에서는 수동 처리 과정에서의 오류를 최소화하고자 수동 주석 과정 시 동일 데이터셋에 대해 작업 후 Fleiss' kappa(κ)를 측정하는 단계를 거쳤고, SPPMI 실험 시에는 숙련된 전문의의 의견이 합치하는 경우에만 seed keyword로 선정하였다.

마지막으로, 제안하는 모델의 전처리 및 키워드 추출 단계에서 한국어와 영어 모두 부정어 처리를 더욱 정교하게 하고, 온톨로지를 고려한 외부 자연어처리 리소스를 사용하는 방법을 고려하면 성능 향상을 꾀할 수 있을 것이다.

5. 추후 연구에 대한 제언

국내에서 생산된 임상기록을 활용하여 자연어처리 연구를 수행할 때 가장 중요하게 고려할 것은, 한국어와 영어가 혼용되고 있다는 점이다. 또한 임상기록이 의료진 별, 진료과 별, 기관 별로 다양할 수밖에 없고, 오탈자 등 철자오류를 포함할 수 있으며, 다수의 두문자나 단어의 나열 등으로 구성되어 있어 의미 있는 얼개를 추출하기 쉽지 않다. 실제로 본 연구를 진행함에 있어서 가장 어려웠던 부분 또한 한영 혼용 임상기록의 전처리였다. 흡연 및 음주 상태 뿐 아니라 문서 내 음주 정보를 일일이 주석작업 하는 과정에 있어서 각 annotator 별로 합치된 레이블 결과를 도출해내는 데까지 최소 3회 이상의 반복 작업을 수행하였고 최소 3개월 이상의 시간과 의료 용어에 익숙한 5명 이상의 annotator의 인력이 필요하였다.

한국어 또는 영어로 구성된 의료 텍스트 데이터에 대한 대규모 구축사업은 진행되고 있으나, 실제 임상 현장에서 생성되는 한국어와 영어가 혼용된 텍스트의 대규모 구축이나 연구는 거의 진행되고 있지 않다. 따라서 텍스트 전처리 과정에서 한국어와 영문을 따로 처리할 수밖에 없어 정보의 손실을 유발할 수 있으며, 기존의 한글 형태소 분석기는 의료영역에 특화되어 있는 것이 아니므로, 효과적인 처리가 어려울 수 있다. KR BERT[87]나 최근 개발된 KM BERT[88] 등 한국어를

지원하는 자연어처리 모델이 있으나, pre-train 과정에서 한국어만을 사용하기 때문에 한영 혼용 임상기록에서는 적극적 활용이 어려울 수 있다. 임상기록에는 비단 흡연, 음주 정보뿐 아니라 방대하고 중요한 임상 정보가 포함되어 있으므로 이를 활용하여 한영 혼재 텍스트를 기반으로 한 새로운 corpus 개발이 필요할 것이다.

본 연구에서 제안하는 방법은, 임상 진료 현장과 연구 현장에서 다음과 같이 활용해볼 수 있다. 먼저, 임상 진료현장에서 이전에 기록되어 저장 되어있는 문서를 자동으로 분류하여, 핵심 정보를 진료 전에 미리 의료진에게 제공하여, 환자에게 필요한 처치를 수행하는데 도움을 줄 수 있을 것이다. 본 연구에서 개발한 알고리즘을 병원정보시스템에 적용하여, 환자 상태를 미리 알람을 주고 필요한 처치를 하게 할 수 있다. 진료 시 반드시 문진이나 진찰을 하여야 하는 경우나, 의료보험 급여 적용을 위한 필수 항목이 기입되어 있는지 손쉽게 검토할 수 있다. 연구에서는, 원하는 정보를 수기로 추출하는 대신 자동으로 추출하여 활용할 수 있을 것이다. 비단 흡연과 음주 정보뿐 아니라 임상기록에는 현재병력, 과거력, 가족력, 신체검진, 약물복용 등 다양한 정보가 존재하므로, 본 연구를 응용하여 선택적인 정보 추출 및 분서 분류를 통해 손쉽게 정보를 획득할 수 있다. 더불어 본 연구의 향후 확장을 위하여 제안하는 방법을 통한 정보 확보 시의 인력, 시간 및 비용 절감과 같은 유용성 평가가 필요할 것이다.

제 5 장 결론

본 연구에서는 전자의무기록 내 한영 혼용 자유진술문으로 구성된 임상기록에서 흡연 음주 정보를 추출하고 문서를 분류하는 알고리즘을 제안하였다. 흡연 정보 추출 및 분류를 위하여 unsupervised learning의 한 방법인 SPPMI를 제안하였으며, 대규모 사전 데이터셋이나 대규모의 자연어처리 플랫폼을 사용하지 않아도, 잠재적인 성능을 보여주었다. 음주 정보 추출 및 분류에는 transformer 기반 multilingual BERT를 활용하였으며 문서를 분류하기 위하여 음주 종류, 잔 수와 같은 보조 정보를 추가하여 성능 고도화를 진행하였다.

본 연구는 이종언어로 구성된 비정형 임상기록에서 의미 있는 키워드를 추출하고 문서를 분류할 수 있는 가능성을 확인한 거의 최초의 연구이다. 영어가 모국어가 아닌 국가는 의학 온톨로지 가용성이 제한되어 표준화가 쉽지 않은 것이 현실이다. 본 연구에서 제안하는 알고리즘은 비영어권 국가의 온톨로지 가용성 한계 극복, 정보 획득에 소모되는 인력과 비용 절감, 임상기록 내 유의미한 정보를 편리하게 확보하는데 기여할 수 있을 것이다.

참고 문헌

1. Doll R, Hill AB. Mortality in relation to smoking: ten years' observations of British doctors. British medical journal 1964;1(5395):1399.
2. 통계청. 현재 흡연율 추이. (https://kosis.kr/statHtml/statHtml.do?orgId=177&tblId=DT_11702_N001&conn_path=I3).
3. 통계청. 고위험음주율 추이. (https://kosis.kr/statHtml/statHtml.do?orgId=177&tblId=DT_11702_N018&conn_path=I3).
4. 통계청. 월간 폭음률 추이. (https://kosis.kr/statHtml/statHtml.do?orgId=177&tblId=DT_11702_N234&conn_path=I3).
5. Yun YH, Jung KW, Bae J-M, et al. Cigarette smoking and cancer incidence risk in adult men: National Health Insurance Corporation Study. Cancer detection and prevention 2005;29(1):15-24.
6. 질병관리청. 흡연으로 인한 국내 사망자 수와 사회경제적 비용 ('19년 기준)2022.
7. MacMahon S. Alcohol consumption and hypertension. Hypertension 1987;9(2):111-121.
8. Blot WJ. Alcohol and cancer. Cancer research 1992;52(7_Supplement):2119s-2123s.
9. Diamond I, Messing RO. Neurologic effects of alcoholism. Western Journal of Medicine 1994;161(3):279.
10. Klatsky AL. Alcohol, cardiovascular diseases and diabetes mellitus. Pharmacological research 2007;55(3):237-247.
11. Rehm J, Taylor B, Mohapatra S, et al. Alcohol as a risk factor for liver cirrhosis: a systematic review and meta-analysis. Drug and alcohol review 2010;29(4):437-445.
12. Boffetta P, Hashibe M. Alcohol and cancer. The lancet oncology 2006;7(2):149-156.
13. Lieber CS. Alcohol, liver, and nutrition. Journal of the American College of Nutrition 1991;10(6):602-632.
14. Roehrs T, Roth T. Sleep, sleepiness, sleep disorders and alcohol use and abuse. Sleep medicine reviews 2001;5(4):287-297.
15. Boden JM, Fergusson DM. Alcohol and depression. Addiction 2011;106(5):906-914.
16. Cherpitel CJ, Borges GL, Wilcox HC. Acute alcohol use and suicidal behavior: a review of the literature. Alcoholism: clinical and experimental research 2004;28:18S-28S.
17. Skog OJ. Alcohol consumption and mortality rates from traffic accidents, accidental falls, and other accidents in 14 European countries. Addiction 2001;96(1s1):49-58.
18. Katikireddi SV, Whitley E, Lewsey J, Gray L, Leyland AH. Socioeconomic status as an effect modifier of alcohol consumption

and harm: analysis of linked cohort data. The Lancet Public Health 2017;2(6):e267-e276.
19. Noble N, Paul C, Turon H, Oldmeadow C. Which modifiable health risk behaviours are related? A systematic review of the clustering of Smoking, Nutrition, Alcohol and Physical activity ('SNAP') health risk factors. Preventive medicine 2015;81:16-41.
20. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? Journal of biomedical informatics 2009;42(5):760-772.
21. Liddy ED. Natural language processing. 2001.
22. Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. Natural language processing in medicine: a review. Trends in Anaesthesia and Critical Care 2021;38:4-9.
23. Kiela D, Hill F, Clark S. Specializing Word Embeddings for Similarity or Relatedness. EMNLP2015:2044-2048.
24. Liu Y, Liu Z, Chua T-S, Sun M. Topical word embeddings. Proceedings of the AAAI Conference on Artificial Intelligence2015.
25. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:13013781 2013.
26. Bodenreider O. The Unified Medical Language System Integrating Biomedical Terminology. 2005.
27. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association 2010;17(5):507-513.
28. Zhang S, Elhadad N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. Journal of biomedical informatics 2013;46(6):1088-1098.
29. Mishra R, Bian J, Fiszman M, et al. Text summarization in the biomedical domain: a systematic review of recent research. Journal of biomedical informatics 2014;52:457-467.
30. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. Journal of biomedical informatics 2018;77:34-49.
31. Jonnagaddala J, Dai H-J, Ray P, Liaw S-T. A preliminary study on automatic identification of patient smoking status in unstructured electronic health records. Proceedings of BioNLP 152015:147-151.
32. Shoenbill K, Song Y, Gress L, Johnson H, Smith M, Mendonca EA. Natural language processing of lifestyle modification documentation. Health informatics journal 2020;26(1):388-405.
33. Wang Y, Chen ES, Pakhomov S, et al. Automated extraction of substance use information from clinical texts. AMIA Annual Symposium Proceedings: American Medical Informatics Association; 2015:2121.
34. Lix L, Munakala SN, Singer A. Automated classification of alcohol use by text mining of electronic medical records. Online Journal of Public Health Informatics 2017;9(1).

35. Alzoubi H, Ramzan N, Alzubi R, Mesbahi E. An automated system for identifying alcohol use status from clinical text. 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE): IEEE; 2018:41-46.
36. Afshar M, Phillips A, Karnik N, et al. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. Journal of the American Medical Informatics Association 2019;26(3):254-261.
37. To D, Sharma B, Karnik N, Joyce C, Dligach D, Afshar M. Validation of an alcohol misuse classifier in hospitalized patients. Alcohol 2020;84:49-55.
38. Fraile-Navarro D, Azcoaga-Lorenzo A, Agrawal U, et al. Development of an algorithm to classify primary care electronic health records of alcohol consumption: experience using data linkage from UK Biobank and primary care electronic health data sources. BMJ open 2022;12(2):e054376.
39. Church K, Hanks P. Word association norms, mutual information, and lexicography. Computational linguistics 1990;16(1):22-29.
40. Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings. Transactions of the association for computational linguistics 2015;3:211-225.
41. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems 2013;26.
42. Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization. Advances in neural information processing systems 2014;27.
43. Wang P, Xu B, Xu J, Tian G, Liu C-L, Hao H. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing 2016;174:806-814.
44. Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. Journal of the American Medical Informatics Association 2008;15(1):14-24.
45. Cohen AM. Five-way smoking status classification using text hot-spot identification and error-correcting output codes. Journal of the American Medical Informatics Association 2008;15(1):32-35.
46. Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U. Identifying smokers with a medical extraction system. Journal of the American Medical Informatics Association 2008;15(1):36-39.
47. Golden SE, Hooker ER, Shull S, et al. Validity of Veterans Health Administration structured data to determine accurate smoking status. Health Informatics Journal 2020;26(3):1507-1515.
48. Groenhof TKJ, Koers LR, Blasse E, et al. Data mining information from electronic health records produced high yield and accuracy for current smoking status. Journal of clinical epidemiology

2020;118:100-106.
49. De Silva L, Ginter T, Forbush T, et al. Extraction and quantification of pack-years and classification of smoker information in semi-structured Medical Records. Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA2011.
50. Figueroa RL, Soto DA, Pino EJ. Identifying and extracting patient smoking status information from clinical narrative texts in Spanish. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society: IEEE; 2014:2710-2713.
51. Patel J, Siddiqui Z, Krishnan A, Thyvalikakath TP. Leveraging electronic dental record data to classify patients based on their smoking intensity. Methods of information in medicine 2018;57(05/06):253-260.
52. Caccamisi A, Jørgensen L, Dalianis H, Rosenlund M. Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records. Upsala journal of medical sciences 2020;125(4):316-324.
53. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American society for information science 1990;41(6):391-407.
54. Hofmann T. Probabilistic latent semantic indexing. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval1999:50-57.
55. Blei DM, Ng AY, Jordan M. Mar. 2003. Latent dirichlet allocation. J Mach Learn Res 2003;3(4-5):993-1022.
56. Matsuo Y, Ishizuka M. Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 2004;13(01):157-169.
57. HaCohen-Kerner Y, Gross Z, Masa A. Automatic extraction and learning of keyphrases from scientific articles. Computational Linguistics and Intelligent Text Processing: 6th International Conference, CICLing 2005, Mexico City, Mexico, February 13-19, 2005. Proceedings 6: Springer; 2005:657-669.
58. Zhu Y, Kiros R, Zemel R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. Proceedings of the IEEE international conference on computer vision2015:19-27.
59. Wenzek G, Lachaux M-A, Conneau A, et al. CCNet: Extracting high quality monolingual datasets from web crawl data. arXiv preprint arXiv:191100359 2019.
60. Gururangan S, Marasović A, Swayamdipta S, et al. Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint arXiv:200410964 2020.
61. Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:190711692 2019.
62. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J.

Deep learning--based text classification: a comprehensive review. ACM computing surveys (CSUR) 2021;54(3):1-40.
63. Conneau A, Lample G, Rinott R, et al. XNLI: Evaluating cross-lingual sentence representations. arXiv preprint arXiv:180905053 2018.
64. Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:190405342 2019.
65. Teo K, Lai KW, Yong CW, Pingguan-Murphy B, Chuah JH, Tee CAT. Prediction of Hospital Readmission Combining Rule-based and Machine Learning Model. 2020 International Computer Symposium (ICS): IEEE; 2020:352-355.
66. Mohammadi R, Jain S, Namin AT, et al. Predicting unplanned readmissions following a hip or knee arthroplasty: retrospective observational study. JMIR medical informatics 2020;8(11):e19761.
67. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ digital medicine 2021;4(1):86.
68. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. Journal of the American Medical Informatics Association 2019;26(11):1297-1304.
69. Kim HK, Choi SW, Bae YS, et al. MARIE: A Context-Aware Term Mapping with String Matching and Embedding Vectors. Applied Sciences 2020;10(21):7831.
70. Ji Z, Wei Q, Xu H. Bert-based ranking for biomedical entity normalization. AMIA Summits on Translational Science Proceedings 2020;2020:269.
71. Abuse NIoA, Alcoholism. Helping Patients Who Drink Too Much, A Clinician's Guide2007.
72. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological bulletin 1971;76(5):378.
73. Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:191102116 2019.
74. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems 2017;30.
75. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. arXiv preprint arXiv:150807909 2015.
76. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog 2019;1(8):9.
77. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805 2018.
78. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:171105101 2017.
79. Lin Y, Sharma B, Thompson HM, et al. External validation of a

machine learning classifier to identify unhealthy alcohol use in hospitalized patients. Addiction 2022;117(4):925-933.
80. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory1992:144-152.
81. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13, 2016; San Francisco, California, USA, 785-794. 2016.
82. Breiman L. Random forests. Machine learning 2001;45:5-32.
83. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234-1240.
84. Kumar A, Narapareddy VT, Gupta P, Srikanth VA, Neti LBM, Malapati A. Adversarial and auxiliary features-aware bert for sarcasm detection. Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)2021:163-170.
85. Balagopalan A, Novikova J. Augmenting BERT Carefully with Underrepresented Linguistic Features. arXiv preprint arXiv:201106153 2020.
86. De Bruyne L, De Clercq O, Hoste V. Emotional RobBERT and insensitive BERTje: combining transformers and affect lexica for Dutch emotion detection. Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), held in conjunction with EACL 2021: Association for Computational Linguistics; 2021:257-263.
87. Lee S, Jang H, Baik Y, Park S, Shin H. Kr-bert: A small-scale korean-specific language model. arXiv preprint arXiv:200803979 2020.
88. Kim Y, Kim J-H, Lee JM, et al. A pre-trained BERT for Korean medical natural language processing. Scientific Reports 2022;12(1):1-10.

Abstract

Natural Language Processing Algorithm for Extraction and Classification of Smoking and Drinking Information from Bilingual Clinical Notes

Ye Seul Bae

Interdisciplinary Program in Medical Informatics

The Graduate School

Seoul National University College of Medicine

Smoking and alcohol consumption are important variable for clinical research, but there are few studies regarding automatic obtainment of classification from unstructured bilingual clinical notes. The study aims to develop an algorithm to classify smoking and alcohol consumption status based on unstructured clinical notes using natural language processing (NLP).

A total of 4,711 clinical notes were used for smoking information extraction and classification. Acronyms were normalized using Soynlp, a Python package. Each group in the record was vectorized

using Shifted Positive Pointwise Mutual Information (SPPMI), and smoking status was classified by calculating cosine similarity. Compared to other methods such as word co-occurrence, pointwise mutual information (PMI), and normalized pointwise mutual information (NPMI), this study improved keyword extraction accuracy by more than 20%. Additionally, the proposed SPPMI method for extracting smoking-related keywords in documents showed the highest precision and achieved a high accuracy of 91.49% in classifying smoking status.

For drinking-related information extraction and classification, a total of 4,996 clinical records were used. The Multilingual transformer was fine-tuned and learned. The proposed cross-lingual language model - robustly optimized BERT pretraining approach (XLM-RoBERTa) model showed high keyword extraction accuracy (macro F1-score 84.7%) compared to other methods and demonstrated excellent performance with improvements in precision and recall, reaching 78.23% and 95.33%, respectively. Furthermore, when additional information about the type of alcohol and alcohol consumption pattern was included in the learning process, the ability to classify documents for drinking information improved by 9.87% in multilingual BERT and 11.7% in XLM-RoBERTa.

This study shows the potential of SPPMI and transformer in classifying smoking status and alcohol consumption from bilingual

clinical notes. Moreover, it is the first study to extract meaningful information and classify documents from clinical records consisting of free text written in both Korean and English.

In the future, the proposed method can be used to capture various information in clinical notes as well as smoking and drinking information. Future validation studies with clinical notes generated by various departments and medical institutions will be necessary.

Keywords : Natural language processing, Smoking, Drinking, Clinical notes, Bilingual language processing, Transformer

Student Number : 2021-31273