The Journal of the Convergence on Culture Technology (JCCT)
Vol. 10, No. 3, pp.391-396, May 31, 2024. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2024.10.3.391
JCCT 2024-5-45

헬스케어 분야 빅데이터 분석을 위한 개체명 사전 구축에 새로운 역 N-Gram 적용 연구

A Study on Applying Novel Reverse N-Gram for Construction of Natural Language Processing Dictionary for Healthcare Big Data Analysis

이경현*, 백락준**, 김우수***

KyungHyun Lee*, RackJune Baek, WooSu Kim*****

요 약 본 연구에서는 헬스케어 분야에 특화된 개체명 사전을 구축하기 위해 기존 N-Gram 방식의 한계를 극복하고 성능을 향상하게 시키기 위해 새로운 역 N-Gram 방식을 제안하였다. 제안된 역 N-Gram 방식은 헬스케어 관련 빅데이터의 복잡한 언어적 특성을 더 정밀하게 분석하고 처리할 수 있다. 제안된 방식의 효율성 검증을 위해 매년 1월에 개최되는 소비자 가전 전시회(Consumer Electronics Show: CES) 기간 동안 발표된 헬스케어 및 디지털 헬스케어 관련 빅데이터를 수집하기 위하여 뉴스를 대상으로 2010년 1월 1일부터 31일, 그리고 2024년 1월 1일부터 31일까지 언급된 2,185건의 뉴스 제목 및 요약문을 파이썬 프로그래밍언어로 새로운 역 N-Gram 방식을 구현하여 전처리한 결과, 헬스케어 분야에서의 자연어 처리를 위한 사전이 안정적으로 구축되었음을 확인할 수 있었다.

주요어 : 자연어 처리, 헬스케어, 역 N-Gram, 빅데이터, 개체명 사전

Abstract This study proposes a novel reverse N-Gram approach to overcome the limitations of traditional N-Gram methods and enhance performance in building an entity dictionary specialized for the healthcare sector. The proposed reverse N-Gram technique allows for more precise analysis and processing of the complex linguistic features of healthcare-related big data. To verify the efficiency of the proposed method, big data on healthcare and digital health announced during the Consumer Electronics Show (CES) held each January was collected. Using the Python programming language, 2,185 news titles and summaries mentioned from January 1 to 31 in 2010 and from January 1 to 31 in 2024 were preprocessed with the new reverse N-Gram method. This resulted in the stable construction of a dictionary for natural language processing in the healthcare field.

Key words : NLP, Healthcare, Reverse N-Gram, Bigdata, NLP Dictionary

Ⅰ. 서 론

최근 헬스케어(Healthcare) 분야는 인구 증가 및 고령화 사회에 맞춰 빠르게 성장하고 있으며, 이 과정에

*정회원, 한국공학대학교 IT반도체융합공학과 박사과정 (제1저자)
**정회원, 가톨릭관동대학교 책임연구원 (공동저자)
***정회원, 한국공학대학교 융합기술에너지 대학원 교수 (교신저자)
접수일: 2024년 3월 12일, 수정완료일: 2024년 4월 22일
게재확정일: 2024년 4월 30일

Received: March 12, 2024 / Revised: April 22, 2024
Accepted: April 30, 2024
***Corresponding Author: kws@tukorea.ac.kr
Graduate school of Convergence Technology and Energy, Tech Univ of Korea

서 생성되는 대량의 데이터를 효과적으로 분석하고 활용하는 것이 중요한 과제로 부상하고 있다.

국내의 대표적인 의료기관인 서울대학교병원의 경우 2009년 기준으로 전자의무기록 시스템 (Electronic Medical Record system, EMR), 처방전달 시스템(Order Communication System, OCS)에 저장된 데이터는 3.4TB이며, 매년 360GB씩 증가하고 있고, 영상정보 시스템(Picture Archiving and Communication System, PACS)은 105TB이고, 매년 32.4TB씩 증가하고 있다고 발표하였다.[1]

이처럼 헬스케어 분야에서 생성되는 데이터는 그 양뿐만 아니라 다양성에서도 급증하고 있으며, 이러한 데이터의 효과적인 분석은 질병 예측, 환자 관리, 치료 방법의 개인화와 같은 혁신적인 의료서비스 제공에 핵심적인 역할을 한다.

수많은 헬스케어 데이터 중에서도 뉴스 기사, 의료 기록, 임상 연구 보고서 등은 헬스케어 전문가들에게 중요한 정보를 제공하지만, 그 내용을 효율적으로 분석하고 활용하기 위해서는 데이터를 정형화하고 특징을 추출할 수 있는 전처리 과정이 필수적이다. 이 과정에서 자연어 처리(Natural Language Process, NLP) 기술은 중요한 역할을 하며, 헬스케어 분야에 특화된 개체명 사전의 구축은 의미 분석의 정확도를 높이는 데 중요한 요소로 자리 잡고 있다.

2022 헬스케어 인공지능 설문조사(2022 Ai Health care Survey)는 전 세계 41개국의 디지털 헬스케어 전문가 321명을 대상으로 시행한 조사에서 새로운 경향의 하나로 텍스트 활용 증가를 제시. 위 조사에서 응답자의 43%는 올해 새로 도입할 기술로 자연어 처리를 선택했으며, 아래 표1은 동일한 키워드로 검색된 국내 논문의 최근 5년간의 경향을 보여준다.[2]

표 1. 키워드 분석 결과표[2]
Table 1. Keyword analysis result table

구분	2017	2018	2019	2020	2021	평균
'헬스케어'+'인공지능'	151	165	199	499	259	254.6
'헬스케어'+'자연어처리'	4	7	8	13	21	10.6

이는 텍스트 데이터의 활용 증가가 새로운 트렌드로 자리를 잡고 있음을 시사하고 있으며, 헬스케어 분야뿐만 아니라 다양한 분야에서 자연어 처리와 관련된 많은 연구가 선행되었다.

▸ 이성직 등(2009년)은 뉴스 문서 집합 전체 범위에서 키워드를 추출하기 위해 6가지의 수정된 TF-IDF(term frequency-inverse document frequency) 가중치 모델과 이를 통해 얻은 키워드 집합을 한층 더 개선하기 위해 분야별 후보 키워드 집합을 통계적으로 교차 비교[3].

▸ 최봉준 등(2017년)은 데이터를 수집하여 저장하고 분석에 이르기까지 오랜 시간이 소모되는 점을 해결하기 위해 실시간 이슈 탐지를 위한 일반-급상승 단어 사전 생성 및 매칭 기법을 제안.[4]

▸ 김규리 등(2020년)은 SNS에 게시된 글의 내용을 통해 사용자의 우울함을 검출하는 기계학습 기반 감성 분석 시스템을 제안[5]

▸ 홍힘찬 (2021년)은 국방 분야 빅데이터 자연어 분석에서 정확성을 높이는 방식을 제시함으로써 효율적이면서도 효과적인 자연어 분석 방법을 제시[6].

▸ 강수연등 (2023년)은 거대 언어 모델의 장점을 활용하여, 한국어의 교착어 특성을 고려한 형태소 정보 기반 Few-Shot 프롬프트 방식을 통한 헬스케어 도메인의 개체명 인식 방법을 제안.[7]

▸ 손현곤 등(2023년)은 의료 문진과 상담 내용을 자동으로 추출, 요약하여 지식화하는 음성인식과 자연어 처리 딥러닝을 통한 득 시스템을 제안.[8]

위의 선행연구처럼 다양한 분야에서 많은 연구가 진행되고 있으며, 본 연구에서는 헬스케어 분야의 개체명 사전 구축에 있어 기존 방식에서 한 단계 더 나아가, 성능을 향상시킨 새로운 역 N-Gram(Reverse N-Gram) 방식을 제안한다. 제안된 방식을 이용하여 헬스케어에 대하여 언급된 뉴스를 분석하여, 뉴스 제목과 요약문에 대한 역 N-Gram 방식의 전처리를 통해, 헬스케어 분야 빅데이터 분석용 개체명 처리 사전을 구축하였다.

헬스케어 분야의 빅데이터 분석에 있어, 새로운 역 N-Gram 방식이 기존 자연어 처리 방법론에 비대 어떤 성능 향상을 보이는지 연구하고자 한다.

본 연구를 통해 헬스케어 분야의 빅데이터 분석과 자연어 처리 기술의 발전에 기여하고자 한다.

II. 역(Reverse) N-Gram

기존의 N-Gram은 카운트에 기반한 통계적 접근을 사용하는 SLM(Statistical Language Model)의 일종으로 주어진 텍스트에서 N개의 음절 또는 단어를 연쇄적으로 분류해 N개의 단어 뭉치 단위로 끊어서 이를 하나의

토큰으로 간주하여 처리한다. N=1일 때 Unigram, 2일 때는 Bigram, 3일째는 Trigram 식으로 불린다. 키워드 추출이 목적이기 때문에 유의미한 N개의 단어를 추출한다.[9]

예를 들어, "I love natural language processing"이라는 문장에서 N-Gram을 적용하면 표2와 같이 Uni-Gram에서 Quad-Gram까지의 4단계의 단어 추출을 나타낼 수 있다.

표 2. 영어 문장 N-Gram 적용 예시
Table 2. Example of N-Gram application in English sentence

N-Gram	적용 예시
Uni-Gram	"I", "love", "natural", "language", "processing"
Bi-Gram	"I love", "love natural", "natural language", "language processing"
Tri-Gram	"I love natural", "love natural language", "natural language processing"
Quint-Gram	"I love natural language", "love natural language processing"

또한 한국어 문장의 경우 예를 들면 "나는 자연어 처리를 좋아해"라는 한국어 문장을 N-Gram을 적용하면 표3과 같이 나타낼 수 있다.

표 3. 한국어 문장 N-Gram 적용 예시
Table 3. Example of N-Gram application in Korean sentence

N-Gram	적용 예시
Uni-Gram	"나는", "자연어" "처리를", "좋아해"
Bi-Gram	"나는 자연어" "자연어 처리를", "처리를 좋아해"
Tri-Gram	"나는 자연어 처리를" "자연어 처리를 좋아해"
Quint-Gram	"나는 자연어 처리를 좋아해"

이러한 N-Gram은 문장 내 단어 간의 관계를 파악하여 언어 모델링을 수행하는 단어 시퀀스 모델링이 가능하다는 장점이 있고, 이를 통해 다음에 나올 단어를 예측하거나 문장의 유사성을 측정하는 등의 작업을 수행할 수 있다. 또한 N-Gram은 단어들을 간단한 형태로 표현하기 때문에 구현이 비교적 간단하고, 대규모 데이터에 대한 처리 등 효율적인 구현이 가능하다.

하지만 대규모 데이터에서는 많은 수의 N-Gram이 출현하지 않을 가능성이 커지는데, 이는 모델의 일반화 성능을 저하할 수 있다.

본 연구에서 제시하고자 하는 역 N-Gram 방식은 자연어 처리에서 텍스트를 처리하고 분석하기 위한 방법 중 하나로, 기존의 N-Gram 방식과는 반대로 단어를 뒤부터 잘라가며 추출하여 분석하는 방식이다.

예를 들어 "아버지께서는"이라는 문자열을 기존의 N-Gram 방식으로 나타내면 표5와 같이 적용하여 추출할 수 있고, 제안하는 새로운 역 N-그램 방식을 적용하면 표6와 같이 표현할 수 있다.

표 5. 한글 문자열 N-Gram 적용 예시
Table 5. Example of application of Korean string N-Gram

N-Gram	적용 예시
Uni-Gram	"아", "버", "지", "께", "서", "는"
Bi-Gram	"아버", "버지", "지께", "께서", "서는"
Tri-Gram	"아버지", "버지께", "지께서", "께서는"
Quint-Gram	"아버지께", "버지께서", "지께서는"
quintal	"아버지께서", "버지께서는"
se	"아버지께서는"

표 6. 한글 문자열 역 N-Gram 적용 예시
Table 6. Example of Korean string Reverse N-Gram application

역 N-Gram	적용 예시
역 Zero-Gram	"아버지께서는"
역 Uni-Gram	"아버지께서"
역 Bi-Gram	"아버지께"
역 Tri-Gram	"아버지"

제안하는 역 N-그램은 위에서 언급한 N-그램의 장점은 살리고, 단점을 보완하기 위해 문자열을 뒤에서부터 앞으로 N개씩 문자를 잘라가며 처리하는 방식이다. 표4은 기존의 N-Gram과 역 N-Gram 개체명 인식 비교

표 4. N-Gram과 역 N-Gram 개체명 인식 비교표
Table 4. Comparison table of N-Gram and reverse N-Gram entity name recognition

문자열	N-Gram		처리 수	역 N-Gram		처리 수
아버지가	Uni	"아", "버", "지", "가"	4	Zero	"아버지가"	1
	Bi	"아버", "버지", "지가"	3	Uni	"아버지"	1
	Tri	"아버지", "버지가"	2	Bi	-	-
	quintal	"아버지가"	1	Tri	-	-
합계			1/10	합계		1/2
아버지께서는	Uni	"아", "버", "지", "께", "서", "는"	6	Zero	"아버지께서는"	1
	Bi	"아버", "버지", "지께", "께서", "서는"	5	Uni	"아버지께서"	1
	Tri	"아버지", "버지께", "지께서", "께서는"	4	Bi	"아버지께"	1
	quint	"아버지께", "버지께서", "지께서는"	3	Tri	"아버지"	1
	quintal	"아버지께서", "버지께서는"	2	quint	-	-
	se	"아버지께서는"	1	quintal	-	-
합계			1/21	합계		1/4

표로 기존 N-Gram에 비해 문자열의 맨 뒤에서부터 일부 문자를 제거하여 인식하는 과정에서 전처리를 간소화할 수 있으며, 데이터에 노이즈가 많거나 정규화가 필요한 경우, 이러한 방식은 데이터의 변형을 감지하고 처리하는 데 도움이 될 수 있다. 일부 문자가 문자열을 찾는 정확도를 높이는 방법은 주어진 데이터 및 상황에 따라 다를 수 있으나 "레벤슈타인 거리(Levenshtein distance)" 또는 "문자열 유사성(cosine similarity)" 등의 알고리즘을 사용하여 문자열 간의 유사성을 비교하여 가장 유사한 것을 선택하는 방법이 있으며, 이에 따라 21개의 문자열과 4개의 문자열에 대한 정확도는 4개의 문자열이 5배의 정확도를 갖는 것으로 계산된다. 또한 데이터 처리의 효율성 측면에서 사용되는 알고리즘의 특성과 데이터의 크기를 고려해야 하는데, 데이터의 크기와 복잡성이 작은 경우 대부분의 방법이 충분히 효율적일 것으로 판단된다. 데이터 분석을 위해 딥러닝 기반의 모델을 사용하는 경우, 모델의 학습 및 추론에는 데이터의 양에 따라 상당한 계산 비용이 들 수 있는데, 특히 모델이 복잡하고 대규모 데이터 세트로 학습된 경우, 모델의 추론 속도가 느릴 수가 있어 상대적으로 데이터의 양이 적은 역 N-Gram 방식의 처리가 높은 효율을 갖는 것으로 판단된다. 역 N-Gram을 수식으로 표현하면 표7과 같다.

표 7. 역 N-Gram 수식적 표현
Table 7. Reverse N-Gram mathematical expression

수 식	설 명
$S = S_1 + S_2 + \cdots + S_n$	문자 n개의 문자열
$F(S_m) = S - \sum_{i=m}^{n} S_i = \begin{cases} True \\ False \end{cases}$ $F(S_{n+1}) = S$ $m = (n+1, n, \cdots, 2)$	문자열의 맨 뒤 문자부터 차례로 N개의 문자를 제거하면서 개체명(자연어) 사전에 존재하는지 (True 혹은 False)를 확인하고 값을 저장
$S_{REM} = S - F(S_m)$	$F(S_m)$가 True이면 문자열 S에서 등록 문자 $F(S_m)$을 제거한 나머지 문자를 버림 문자로 등록
$S_{NER} = F(S_m)$	개체명(자연어) 사전에 존재하는 문자 지정

표 7의 수식에서 $F(S_{n+1})$을 문자열 S로 조건 처리한 것은 문자열 자체가 개체명으로 인식될 수 있는 경우를 대비하여 조건 처리한 것이다.

$m = (n+1, n, \cdots, 2)$적용 조건은 $m = 1$인 경우에 문자열의 문자가 없어지게 되어 무의미한 루틴으로 처리되는 것을 보완하고자 한 조건이다.

Ⅳ. 개체명 사전 구축

개체명 사전은 빅데이터 분석에서 정보의 정확한 추출, 깊이 있는 분석, 문맥적 이해 및 언어 모델의 성능 향상에 필수적이고, 이를 통해 더 정확하고 유용한 분석 결과를 얻을 수 있다.

헬스케어 관련하여 수집된 텍스트에 대하여 문장에서 문자열을 분리 추출하고 역 N-Gram을 활용하여 개체명 사전을 구축하는 절차를 그림1에 나타내었다.

그림 1. 개체명 사전 구축 절차
Figure 1. Personal name dictionary construction procedure

▸ 텍스트 수집은 인터넷에서 웹 크롤링(web Crawling)과 기존의 말뭉치(Corpus)를 활용하여 분석하고자 하는 텍스트 자료를 수집하였다. 데이터는 헬스케어에 대한 이슈와 동향을 얻으려는 방안으로 소비자 가전 전시회(Consumer Electronics Show:CES)가 개최되는 1월을 기준으로 하여 2010년도 1월 1일~31일, 2024년도 1월 1일~31일 키워드 “헬스케어”와 “디지털 헬스케어”에 대한 뉴스의 제목과 요약 문장을 기초하여 데이터를 수집하였다.

▸ 토큰화(Tokenization)는 텍스트를 문장부호, 대소문자 등을 고려하여 기준을 세우고 문장 또는 단어 단위로 토큰화한다.

▸ 토큰화된 단어들을 이용하여 역 N-Gram을 생성한다. 기존 N-Gram 방식은 앞에서부터 단어를 잘라내기 때문에 단어의 뒷부분에 나타나는 정보가 손실될 수 있으나, 역 N-Gram 방식을 사용한다면 단어의 뒷부분이 더 중요한 경우에도 해당 정보를 보존할 수 있으며, 단어의 뒷부분을 따로 분리하여 처리할 수 있어서 어근 추출이 유용하다.

그림2는 역 N-Gram 모듈 순서도이다.

N-Gram의 N값을 RNG 변수에 대입하여 Word의 길이에 따른 새로운 값 Word_F를 구하여 사전에 등록되어 있는지를 판단하고 RNG 값을 Word 길이의 “-1”까지 진행하여 예비 개체명 사전에 저장하고 NRG의 값이 Word의 길이 값과 동일하다면, 개체명 사전에 등록되어 있지 않은 Word 라면 Word 자체를 예비 개체명 사전에 등록시킨다. 예비 개체명 사전에 저장된 데

이터는 반자동 등록처리 모듈을 통해 헬스케어 개체명 사전에 등록시킨다.

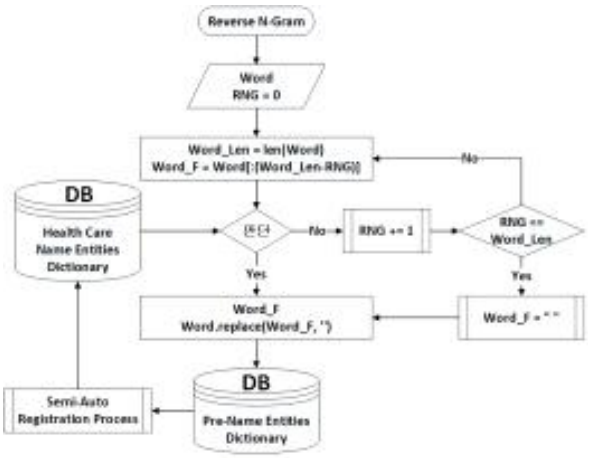

그림 2. 역 N-Gram 모듈 순서도
Figure 2. Reverse N-Gram Module flow chart

역 N-Gram Python 프로그램 코드는 표8과 같다.

표 8. 역 N-Gram Python 프로그램 코드
Table 8. Reverse N-Gram Python Program Code

```
Word_s = Word_Regi(List, DB_Name ,Table_Name)
Word_L = []
for Word in Word_s:
    RNG = 0
    Word_Len = len(Word)
    while True:
        Word_UF = Word[:(Word_Len-RNG)]
Check_Word=NED_Data_Handle.Check_Word_Dict(session
['Field'], Word_UF)
        if Check_Word == "Yes":
            Word_S_L = Word_UF
            Word_L_D = [Word, Word_S_L,
Word.replace (Word_S_L, '')]
            Word_L.append(Word_L_D)
            break
        else:
            RNG += 1
            if RNG == Word_Len:
                Word_S_L = Word
                Word_L_D = [Word, '', Word_S_L]
                Word_L.append(Word_L_D)
                break
```

표 8에서 구현된 파이썬 프로그램 tool로 역 N-Gram의 사용 예를 들어보자면 다음과 같다.

“헬스케어의 사전적 의미를 살펴보면 '넓은 의미로 기존 치료 부문의 의료서비스에, 질병의 예방 및 관리를 포괄하는 전반적 건강관리 사업을 일컫는다'라고 기술되어 있고 좁은 의미로는 '원격 검진이나 방문 건강상담 등의 사업을 지칭한다'라고 정의되어 있다.”라는 수집한 문장에서 단어를 토큰화하여 예비 DB(Database)에 저장한다. 토큰화된 문자열은 다음과 같다.

“헬스케어의”, “사전적”, “의미를”, “살펴보면”, “넓은”, “의미로”, “기존”, “치료”, “부문의”, “의료서비스에”, “질병의”, “예방”, “및”, “관리를”, “포괄하는”, “전반적”, “건강관리”, “사업을”, “일컫는다”, “고”, “기술되어”, “있고”, “좁은”, “의미로는”, “원격”, “검진이나”, “방문”, “건강상담”, “등의”, “사업을”, “지칭한다”, “고”, “정의되어”, “있다”

토큰화된 문자열을 역 N-Gram 처리를 진행하면 다음과 같이 정리되어 개체명 사전에 등록한다.

“헬스케어”, “사전”, “의미”, “치료”, “의료서비스”, “질병”, “예방”, “관리”, “표괄”, “건강관리”, “사업”, “기술”, “의미”, “원격”, “검진”, “방문”, “건강상담”, “사업”, “지칭”, “정의”

위와 같은 과정을 통해 매년 1월에 개최되는 소비자 가전 전시회(Consumer Electronics Show: CES) 기간 동안 발표된 헬스케어 및 디지털 헬스케어 관련 빅데이터를 수집하기 위하여 뉴스를 대상으로 2010년 1월 1일부터 31일, 그리고 2024년 1월 1일부터 31일까지 언급된 2,185건의 뉴스 제목 및 요약문을 역 N-Gram 기반의 개체명 사전을 구축하였다.

구축 결과는 표9과 같이 구축하였으며, 2010년 대비 2024년의 개체명 수가 적은 것은 2010년 이미 구축된 개체명 사전에 등록된 개체명을 등록 개체로 인식하여 추가 등록 작업을 진행하지 않은 결과이다.

표 9. 역 N-Gram 개체어 사전 구축 결과
Table 9. Result of Reverse N-Gram object dictionary construction

	키워드			
	헬스케어		디지털 헬스케어	
기간	뉴스 수	개체 수	뉴스 수	개체 수
2010년 1월 1일~31일	599건	3,713개	103건	1,532개
2014년 1월 1일~31일	749건	1,487개	734건	1,205개
총 뉴스 건	2,185건			
총구축 개체명 수	7,757개			

V. 결 론

본 연구에서 구축한 개체명 사전의 결과를 보면, 2010년 1월 1일부터 31일까지와 2014년 같은 기간 동안의 '헬스케어' 및 '디지털 헬스케어' 관련 뉴스 총수는 2,185건에서 역 N-Gram을 통해 7,757개의 개체어 사전을 구축하였으며, 역 N-Gram 방식을 통한 개체어 사전 구축이 가능하다는 것을 확인하였다. 본 연구 결과는 헬스케어 분야에서의 자연어처리 기술의 발전 가능성을 시사하며, 실제 세계의 이슈와 동향을 파악하는 데에도 중요한 역할을 할 수 있음을 확인하였다. 이러한 방식은 헬스케어 데이터의 복잡성과 다양성을 효과적으로 처리하고, 해당 분야의 빅데이터 분석 및 인사이트 도출에 크게 기여할 것으로 기대한다.

향후 추진 연구에서는 이 사전을 활용하여 다양한 헬스케어 응용 분야에서의 실질적인 적용 가능성과 토픽 모델링(Topic Modeling) 기술을 탐색하고, 역 N-Gram 방식의 확장과 최적화를 통해 자연어 처리 기술의 발전을 이끌어내고자 한다.

References

[1] 신수용, "비정형 헬스케어 데이터 표준화," The Journal of The Korean Institute of Communi cation Sciences, vol. 35, no. 2, pp. 58-64, 2018.

[2] R&D BRIEF "Natural Language Processing in Healthcare" NRF한국연구재단 42호, 2022

[3] Sungjick Lee, Han-joon Kim, "Keyword Extraction from News Corpus using Modified TF-IDF" 한국전자거래학회지 vol.14, no.4, pp. 59-73, 2009

[4] Bongjun Cho, HanJoo Lee, Wooseok Yong, and Won Suk LEE, "A Generation and Matching Method of Normal-Transient Dictionary for Realtime Topic Detection," The Journal of Korean Institute of Next Generation Computing, vol. 13, no. 5, pp. 7-18, 2017.

[5] Kyuri Kim, Jihyun Moon, Uran Oh, "Analysis and Recognition of Depressive Emotion through NLP and Machine Learning ," The Journal of the Convergence on Culture Technology (JCCT), vol.6, no.2, pp.449-454, 2020.

[6] Himchan Hong, "Building a Natural Language Processing Dictionary for Analysing Military Areas' Bigdata," Korean Journal of Military Art and Science, vol. 77, no. 2, pp. 400-415, 2021.

[7] Su-yeon Kang and Gun-woo Kim. "Morpheme -Based Few-Shot Learning with Large Language Models for Korean Healthcare Named Entity Recognition." 한국정보처리학회 학술대회 논문집, vol. 30, no. 2, pp. 428-429, 2023.

[8] Hyeon-kon Son, Gi-hwan Ryu, "Automatic Electronic Medical Record Generation System using Speech Recognition and Natural Language Processing Deep Learning" The Journal of the Convergence on Culture Technology(JCCT), vol.9, no.3, pp.731-736, 2023

[9] Dokyoung Kim and Yu-Seop Kim, "Development of Chinese Media Keyword Analysis System using TF-IDF and N-gram," in 한국정보과학회 학술발표논문집, pp.1432-1434. 2020

[10] Geonwoo ParkO, Seongsik Park, Yoengjin Jang, Kihyoen Choi, Harksoo Kim. "KACTEIL-NER: Named Entity Recognizer Using Deep Learning and Ensemble Technique" Kangwon National University Computer and Communication Engineering pp 324~326, 2017

[11] Jae-Kyun Kim, Chang-Hyun Kim, Min-Ah Cheon, Ho-Min Park, Ho Yoon, Young Nam-Goong, Min-Seok Choi, Jae-Hoon Kim. "Generating Korean NER Corpus using Hidden Markov Model"Korea Maritime and Ocean University, Electronics and Telecommunica tions Research Institute. pp357~361, 2019

[12] Yoon-Shik Tae, Seong-Bae Park, Sang-Jo Lee, and Se-Young Park, "Self-Organizing n-gram Model for Automatic Word Spacing," in 한국정보과학회 언어공학연구회 학술발표 논문집, pp. 125-132. 2006

[13] Dongyoung Lee, "Natural Language Processing Research," in 한국정보과학회 학술발표논문집, pp. 1771-1773. 2018

[14] Sungwook Ko and Hyeryung Jang, "A Study on Improving Practicality for Natural Language Processing Applications Based on a Pre-trained Language Model," in Proceedings of Symposium of the Korean Institute of communications and Information Sciences, pp. 909-910. 2023

※ 이 논문은 2024년 산업통상자원부 산업혁신기반구축사업의 지원에 의하여 연구되었음(P0025775)