종 설

https://doi.org/10.17340/jkna.2021.3.2

뇌졸중 연구에서 자연어처리와 텍스트 마이닝의 적용

김철호

한림대학교 의과대학 춘천성심병원 신경과

An Implementation of Natural Language Processing and Text Mining in Stroke Research

Chulho Kim, MD

Department of Neurology, Chuncheon Sacred Heart Hospital, Hallym University College of Medicine, Chuncheon, Korea

Natural language processing (NLP) is a computerized approach to analyzing text that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. In healthcare field, these NLP techniques are applied in a variety of applications, ranging from evaluating the adequacy of treatment, assessing the presence of the acute illness, and the other clinical decision support. After converting text into computer-readable data through the text preprocessing process, an NLP can extract valuable information using the rule-based algorithm, machine learning, and neural network. We can use NLP to distinguish subtypes of stroke or accurately extract critical clinical information such as severity of stroke and prognosis of patients, etc. If these NLP methods are actively utilized in the future, they will be able to make the most of the electronic health records to enable optimal medical judgment.
J Korean Neurol Assoc 39(3):121-128, 2021

Key Words: Stroke, Natural language processing, Machine learning

서 론

정보기술과 데이터 저장 및 처리기술이 점차 고도화되면서, 의료영역에서도 환자와 관련된 수많은 데이터를 이용하여 환자의 치료를 최적화하거나 진단을 보조할 수 있는 시대가 되었다.[1] 그러나 이러한 정보의 홍수는 반대로 그만큼 환자를 치료하기 위해 확인해야 되는 정보의 양도 많아졌다는 것을 의미한다.[2] 쌓여가는 전자의무기록(electronic medical record, EMR) 데이터 중에서 정형데이터가 차지하는 비중보다는 영상, 시그널 또는 텍스트와 같은 비정형데이터가 차지하는 비중이 월등이 높은데,[3,4] 이렇게 방대한 비정형데이터를 일일이 확인하는 것은 불가능에 가깝다.

자연어처리(natural language process)는 인간의 언어 현상을 컴퓨터와 같은 기계를 이용해서 묘사할 수 있도록 연구하고, 이를 구현할 수 있도록 하는 인공지능의 주요 연구분야의 하나이다.[5,6] 여기서 자연어란 자연적으로 발생하여 사람들의 의사소통에 사용되는 언어로, C, R 또는 파이썬과 같은 컴퓨터 프로그래밍 언어와 같이 사람에 의해 의도적으로 만들어진 인공어(constructed language)에 대비되는 개념이다.[7,8] 즉 자연어는 한글, 영어와 같이 일상적인 의사소통을 위해 자연적으로 만들어진 언어로, 이러한 자연어를 컴퓨터가 이해하고, 기계가 읽을 수 있도록 가공 및 처리하는 일련의 과정을 자연어처리라고 한다.[9] 의료에 있어 이러한 자연어처리 기법은 치료의 적정성에 대한 평가, 질병의 이환 여부 평가 및 임상의사결정보조에 이르는 다양한 응용분야에 적용되고 있다.[10] 의학용어나 EMR 데이터로부터 특정 정보를 추출하는 자연어처리의 능력은 환자의 치료를 최적화하고 인간이 가질 수 있는 의학적 오류를 최소화할 수 있어, 정밀의학을 실현할 수 있는 기술의 핵심이자 가장 활발히 인공지능을 적용할 수 있는 분야이다.

혹자는 텍스트 마이닝과 자연어처리에 대한 정확한 구분을 요구할 수도 있다. 텍스트 마이닝은 텍스트 형태의 비정형데이터로부터 새로운 고급 정보를 이끌어내는 과정을 이야기하며, 이 경우에

Received February 14, 2021 **Revised** June 22, 2021
Accepted June 22, 2021
Address for correspondence: Chulho Kim, MD
Department of Neurology, Chuncheon Sacred Heart Hospital, Hallym University College of Medicine, 77 Sakju-ro, Chuncheon 24253, Korea
Tel: +82-33-240-5255 Fax: +82-33-255-6244
E-mail: gumdol52@naver.com

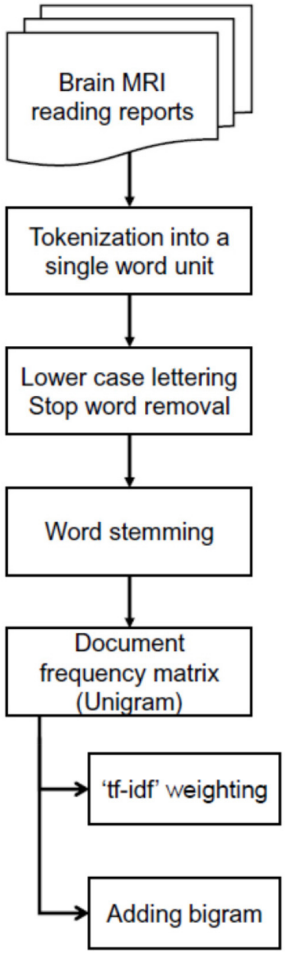
는 문서의 텍스트의 구조를 중요시하지 않는다.[11] 자연어처리의 경우에는 일반적으로 텍스트의 구조를 중요시하며, sentence splitting, part-of-speech tagging 또는 parse tree construction 등의 방법을 사용하여 텍스트로부터 문맥이나 의미를 찾아내는 과정이라고 이해할 수 있다.[12] 그러나 최근의 경우에는 자연어처리 방법을 이용한 텍스트 마이닝이 많이 사용되고 있어, 본문에서는 이러한 구분 없이 자연어처리로 기술하였다. 본 종설에서는 뇌졸중과 관련된 자연어처리 연구들을 정리하여 자연어처리에 대한 개념을 간단히 소개하고, 향후 뇌졸중 연구에서 자연어처리의 전망에 대해 기술하겠다.

본 론

1. 텍스트의 벡터화

1) 토큰과 텍스트벡터를 만들기 위한 자연어 전처리

환자들의 임상기록 및 방사선 보고서와 같이 텍스트로 이루어진 보고서를 기계가 읽도록 만들어 주는 과정을 전처리(preprocessing) 과정이라고 한다.[13] 여기에서 토큰(token)은 텍스트 분석의 기본이 되는 단위로 음절, 단어 구, 문장 등이 토큰의 단위가 될 수 있다. 가장 일반적으로는 한 단어를 토큰으로 하여 텍스트의 모든 내용을



e.g.

1. Multiple small diffusion restrictions in the left middle cerebral artery territory.

Multiple	small	diffusion	restrictions	in	the
left	middle	cerebral	artery	territory	

multiple	small	diffusion	restrictions		
left	middle	cerebral	artery	territory	

multipl	small	diffus	restrict		
left	middl	cerebr	arter	territori	

multipl	small	diffus	restrict	left	middl	cerebr	arter	territori
1	1	1	1	1	1	1	1	1

multipl	small	diffus	restrict	left	middl	cerebr	arter	territori
0.258	0.357	1.339	0.126	0.111	0.663	0.373	1.634	3.183

multipl	small	diffus	restrict	left	middl	cerebr	arter	territori
1	1	1	1	1	1	1	1	1
multipl_small		small_diffus		diffus_restrict		restrict_left		
1		1		1		1		
left_middl		middl_cerebr		cerebral_arter		arter_territori		
1		1		1		1		

Figure 1. Natural language preprocessing flow chart for the corresponding texts. MRI; magnetic resonance imaging, tf-idf; term frequency inverse document frequency.

토큰화하고, 대소문자 통일, 단어 어간추출(word stemming) 및 제외어 제거(stop word removal) 등을 거쳐 매트릭스(matrix) 형태의 벡터 데이터프레임을 형성하는 것이 텍스트 전처리 과정의 일반적인 과정이다(Fig. 1).[14] 여기서 어간추출이라 함은 "artery, arterial, arterio" 등과 같이 같은 의미로 사용되고 있는 단어들의 기원인 "arter"로 토큰을 치환하는 것을 의미한다.[15] 또한 제외어는 텍스트벡터를 만들 때 키워드로 하지 않는 언어들로, 검색용어로 잘 사용하지 않는 "and, is, in, therefore"와 같은 관사, 전치사, 조사, 접속사 등 의미가 없는 단어들을 제거하는 과정이 제외어 제거과정에 해당한다.[16]

2) Concept mapping and rule-based mechanism

Concept mapping 방법은 가공하지 않은 원래의 텍스트로부터 중요한 개념을 뽑아내는 방법으로 MetaMap, MedLEE, cTAKES 등이 많이 사용되고 있다.[17-21] 간략하게는 concept mapping은 전처리된 텍스트로부터 이미 사전에 고유하게 등록되어 있는 중요한 의학용어들을 추출하는 방법을 말한다. MetaMap은 Aronson[19,22]에 의해 고안된 방법으로 Nation Library of Medicine에서 추출된 텍스트를 Unified Medical Language System Metathesaurus에 등록되어 있는 concept unique identifier가 있는 단어들로 매핑하는 방법이다. 이렇게 매핑되어 있는 identifier의 조합과 특별한 컴퓨터 언어학을 이용하여 해당 텍스트로부터 중요한 의미를 뽑아내는 방법 중의 하나이다.

3) 텍스트 임베딩

텍스트 임베딩이란 자연어를 벡터로 변환하는 기법을 말하는 것으로, 이는 텍스트를 벡터공간으로 "끼워넣는다(embed)"라는 의미에

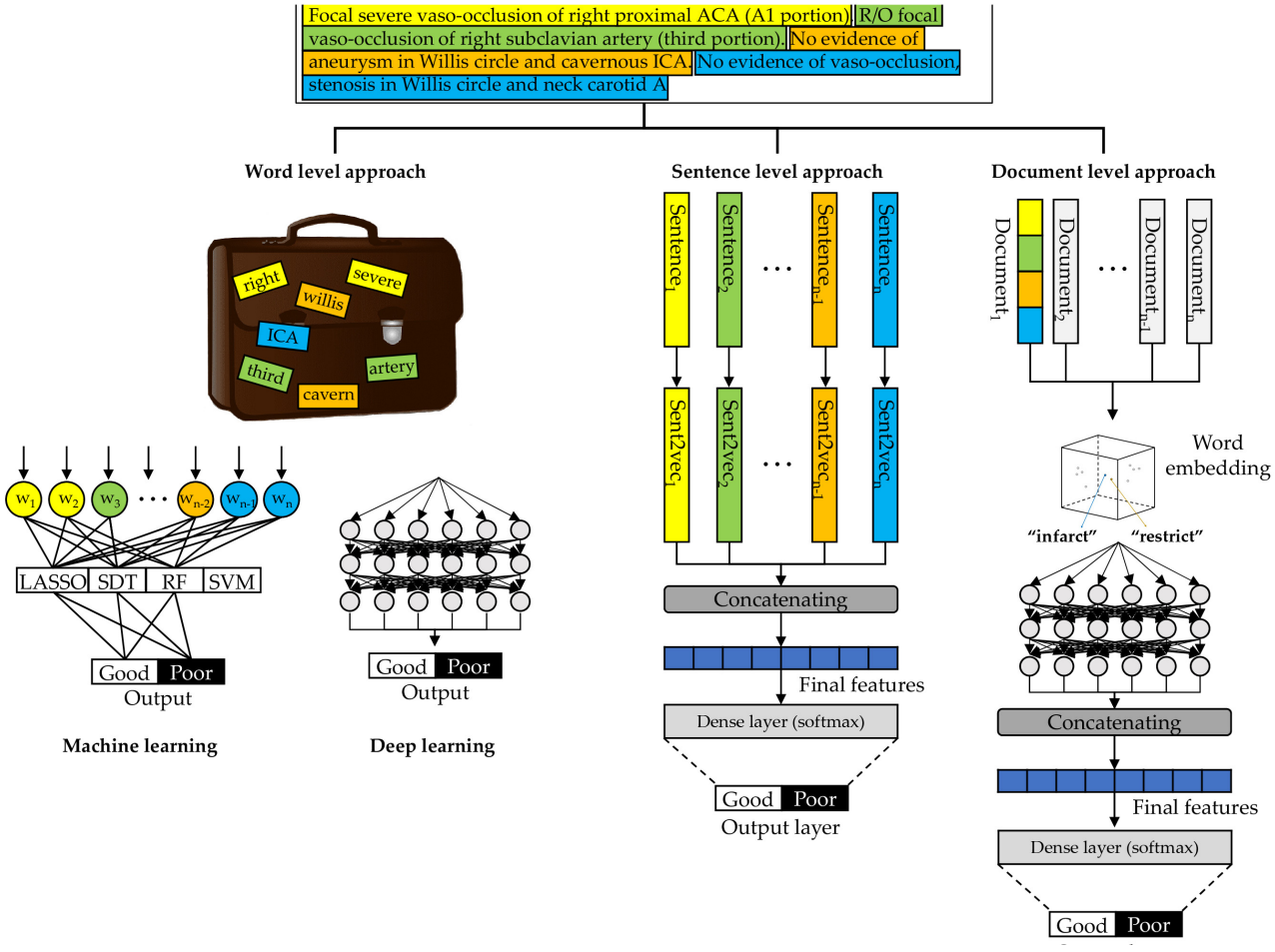


Figure 2. Example natural language processing and machine learning of brain magnetic resonance imaging radiology text reports. ACA; anterior cerebral artery, R/O; rule out, ICA; internal carotid artery, LASSO; least absolute shrinkage and selection operator, SDT; single decision tree, RF; random forest, SVM; support vector machine.

서 유래되었다. 즉 자연어를 기계가 이해할 수 있고 연산과 처리가 가능한 벡터로 바꾸게 되면, 문장의 의미를 계산할 수 있고 사칙연산이 가능하여 컴퓨터에게 의미나 문법적인 정보를 산술적으로 전달할 수 있게 된다.[23] 텍스트 분석의 단위를 말뭉치(corpus)라고 하는데,[24] 이러한 말뭉치의 통계적인 패턴을 단어, 순서, 분포에 따라 임베딩을 할 수 있다. 단어가 중심일 경우에는 "bag-of-words"라는 모델을 사용한다.[25] 여기서 "bag"이란 중복을 제외하지 않은 단어들의 집합을 얘기하고, Fig. 2에서와 같이 해당 말뭉치는 여러 토큰과 해당 토큰의 빈도로 구성된 데이터프레임 형태의 벡터로 만들어지게 된다.[26] 또한 n-gram이라는 것은 토큰들의 순서들까지 같이 고려하여 n개의 토큰을 순서대로 묶어 벡터를 만드는 것을 말한다.[27] 마지막으로 워드 임베딩(word embedding)이라는 방법은 어휘의 단어나 구문이 실제 숫자의 벡터에 매핑되는 자연어처리의 언어 모델링 및 기능 학습기법 중 하나로, 위키피티아 또는 네이버지식인과 같이 대량의 텍스트들의 단어들을 학습하여 수치형 벡터로 만드는 과정을 말한다.[28,29] 이렇게 학습을 하는 과정에는 현재 내가 관심이 있는 문장 또는 단어만을 학습시키는 "task-specific" 워드 임베딩 방법과 함께 앞서 기술한 위키피티아 또는 페이스북 등 잘 알려져 있는 텍스트 집합의 단어나 문장 전체를 학습하는 "pretrained" 워드 임베딩 방법이 있다.[30,31]

2. 자연어처리 분석

해당되는 텍스트로부터 전처리 과정을 거쳐 컴퓨터가 읽을 수 있도록 한 이후, 우리는 자연어처리 과정을 통해 중요한 정보들을 추출할 수 있는데, 이러한 과정에는 크게 3가지 방법이 사용되고 있다. 첫째, 가장 오래 전부터 사용되고 있는 방법인 룰기반알고리즘(rule-based algorithm)은 정규식 표현이나 이미 정해져 있는 문맥자유문법(context free grammar)을 이용하여 패턴을 매칭하거나 비어 있는 결과 컬럼을 채워 넣은 형태로 정보들을 얻게 되는 경우이다.[32] 이러한 방법은 자료들을 학습시키지 않아도 된다는 장점이 있으나, 다른 데이터에 적용하기 위한 일반화에는 한계가 있을 수 있다. 둘째, 고전적인 머신러닝 방법이 이에 해당하는데, 로지스틱 회귀분석, 서포트벡터머신, 의사결정나무(decision tree) 등의 방법과 앙상블 머신러닝 방법인 랜덤 포레스트 및 extreme gradient boosting 등이 대표적인 방법이다.[33,34] 텍스트 데이터프레임을 이용하여 특정 표현형으로 구분하거나(classification), 특정 수치형 데이터를 산술적으로 예측(regression)하는 방법으로 이루어지며, 이를 통해 우리는 대량의 텍스트 데이터로부터 뇌경색 환자를 찾아내거나, 뇌졸중의 아형과 같은 특정 표현형을 찾아낼 수 있다. 세 번째로 뉴럴네트워크 방법은 머신러닝 방법과 유사한 방법이나 뉴런의 형태와 같은 층들을 다양하게 쌓아 벡터들의 가중치들을 종합하여 결과를 얻어내는 방법이다.[35] 이러한 뉴럴네트워크 방법은 단어들의 순서(sequence)를 인식할 수 있는 순서모델을 설계할 수 있어, 최근 자연어처리에서 고전적인 머신러닝 방법에 비해 더 많이 사용되고 있다.[36] 또한 자연어처리 분석에 있어 뉴럴네트워크가 고전적인 머신러닝 방법에 비해 일반적으로 성능이 훨씬 높은 것으로 보고되고 있어, 합성곱신경망(convolutional neural network), 순환신경망(recurrent neural network) 및 encoder-decoder 등의 뉴럴네트워크 방법이 많이 사용되고 있다.[37]

3. 자연어처리 기반의 뇌졸중 연구

저자는 자연어를 이용하여 "PubMed"에 보고된 연구들의 방법 및 결과를 Table에 제시하였다. Elkins 등[38]은 The Northern Manhattan Study 데이터를 이용하여 computed tomography (CT)와 magnetic resonance imaging (MRI) 판독지를 규칙기반 알고리즘으로 분석하였다. 해당 텍스트 데이터에 뇌졸중 병변이 있는 것인지를 예측하기 위한 성능은 area under the receiver operating characteristics (AUROC) 85%로 높지는 않았으나, MedLEE를 이용한 첫 뇌졸중관련 자연어처리 연구로 자동적으로 CT나 MRI 판독지 텍스트를 이용하여 새로운 정보들을 자동적으로 추출해낼 수 있다는 것을 보인 첫 번째 연구이다. 이와 같이, 자연어처리의 적용은 다양한 목적을 위해 사용될 수 있는데, Pons 등[39]은 자연어처리 기반 연구의 목적을 1) 진단서베일런스, 2) 역학 연구를 위한 코호트 구축, 3) 영상검사의 질평가, 4) 임상의사결정지원시스템(clinical decision support system)으로 구분하였다. 진단서베일런스는 특정질환이나 급성 뇌경색과 같은 관심이 있는 특정 소견에 대한 정보를 자동적으로 얻는 것을 말한다. Mowery 등[40]은 경동맥초음파 판독지 및 임상경과 레포트를 PyConText 매핑을 이용하여 의미 있는 경동맥협착의 정보를 정확하게 얻을 수 있는지 조사하여, 민감도 73-88% 및 특이도 84-87%의 예측성능을 보고하였다. 또한 Kim 등[41]은 특정 MRI 판독지 데이터셋으로부터 급성기 뇌경색 환자들을 분류해낼 수 있는 자연어처리 기반의 머신러닝 연구 결과를 제시하였다. 3,204명의 MRI 판독지를 분석함에 있어 "bag-of-word" 모델과 n-gram이 성능에 미치는 결과를 분석하였는데, 단순한 "bag-of-word" 모델이 계산력이 필요한 n-gram 방법과 비교하여 성능이 크게 차이가 나지 않는다는 결과를 보여주었다.[41] 이와 같이 특정질환군을 판독지로부터 찾아낼 수 있는 자연어처리 방법은 코호트 연구나 역학 연구에서 환자군을 추출하는 데 있어 상당한 시간소모를 줄일 수 있다.[42] 자연

어처리 방법을 이용한 영상검사의 질평가에 대한 뇌졸중 관련 연구에 대해 보고는 없었고, 나머지 연구들의 경우 대부분이 임상의사결정지

Table. Studies of natural language processing-based stroke research

Study	Algorithm	NLP preprocessing	Participants	Type of texts	Outcome	Metrics	Cross-validation	External validation
Elkins et al.[38] (2000)	Rule-based mechanism	MedLEE concept mapping	471	CT, MRI reports	Lesion acuity, location	AUROC: 85%	-	No
Mowery et al.[40] (2016)	Rule-based mechanism	PyConText mapping	498	Carotid USG reports and clinical notes	Carotid stenosis detection	Sensitivity: 73-88%, specificity: 84-87%	-	No
Sung et al.[50] (2018)	Rule-based mechanism	MetaMap UMLS concept mapping	78	Clinical notes	Identifying the concept regarding IVT	F1: 99.8	-	No
Kim et al.[41] (2019)	ML	BOW, word embedding	3,204 reports	All brain MRI reports	Phenotype classification	F1: 93.2 (AIS vs. non-AIS)		
Bacchi et al.[47] (2019)	ML & DL	BOW	2,201	TIA referral free texts	CVA TIA vs. non-CVA TIA	AUROC: 81.9	10-fold	No
Alex et al.[44] (2019)	Rule-based mechanism	POS, lemmatization	1,168 reports	Stroke CT, MRI reports	Classification (inter-annotator agreement)	F1: >90.0	-	No
Wheater et al.[45] (2019)	Rule-based mechanism	POS, lemmatization	1,692	Stroke CT, MRI reports	24 phenotype	Ischemic stroke sensitivity (89%), specificity (100%)	-	Yes
Fu et al.[46] (2019)	Rule-based mechanism, ML & DL	MedTagger concept mapping	1,000	CT, MRI reports	SBI, WMH	Sensitivity (92.5), specificity (100%) for SBI	-	No
Garg et al.[48] (2019)	ML	BOW, cTAKES concept mapping	1,091	Radiology report, clinical notes +	TOAST classification	Kappa (0.57)	5-fold	No
Heo et al.[52] (2020)	ML & DL	BOW, word embedding	1,840	AIS brain MRI texts	Outcome classification	AUROC: 80.5%	5-fold	No
Sung et al.[49] (2020)	ML	MetaMap UMLS concept mapping	4,640	AIS clinical notes (admission note only)	OCSP multiclass classification	Accuracy: 0.583	10-fold	No
Kogan et al.[51] (2020)	Rule-based mechanism	CMS mapping	7,149	Physicians note	Maximum NIHSS score	RMSE: 0.45	3-fold	No
Ong et al.[43] (2020)	ML & DL (LSTM)	BOW, word embedding (GloVe)	1,359	Stroke CT, MRI reports	IS, MCA, acuity	AUROC: 0.93-0.98	10-fold	Yes
Li et al.[42] (2021)	ML	BOW with N-grams	32,555	All brain MRI texts in stroke center	AIS detection	F1: 0.74	5-fold	Yes
Guan et al.[54] (2021)	ML	Regular expression	1,598	ICD codes + echocardiography text	CE stroke (vs. non-CE)	AUC: 91.1	5-fold	No

NLP; natural language processing, CT; computed tomography, MRI; magnetic resonance imaging, AUROC; area under the receiver operating characteristic curve, USG; ultrasonography, UMLS: Unified Medical Language System, IVT; intravenous thrombolysis, F1; F1-score (harmonized mean of precision and recall), ML; machine learning, BOW; bag-of-words, AIS; acute ischemic stroke, DL; deep learning, TIA; transient ischemic attack, CVA; cerebrovascular disease, POS; part-of-speech, SBI; silent brain infarct, WMH; white matter hyperintensity, TOAST; Trial of Org 10172 in Acute Stroke Treatment, OCSP; Oxfordshire Community Stroke Project, CMS; The Centers for Medicare & Medicaid Services, NIHSS; National Institutes of Health Stroke Scale, RMSE; root mean square error, LSTM; long-short term memory, GloVe; global vectors for word representation, IS; ischemic stroke, MCA; middle cerebral artery, ICD; International Classification of Diseases, CE; cardioemboli, AUROC; area under the receiver operating characteristic.

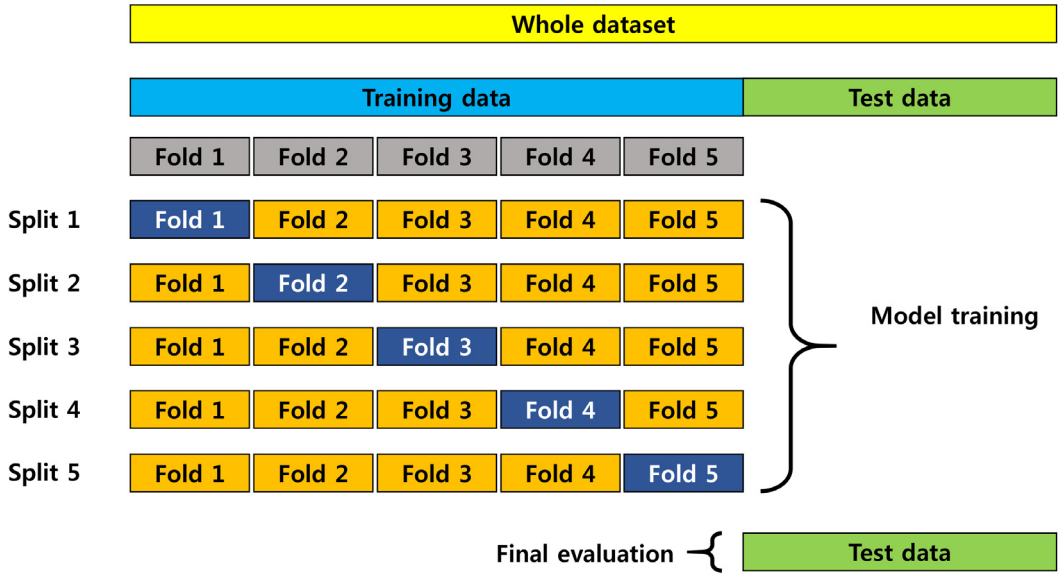


Figure 3. Example of 5-fold cross-validation in machine learning.

원시스템에 대한 연구 결과였다. 이 임상의사결정지원시스템은 환자로부터 얻어진 임상정보를 바탕으로 의료인이 질병을 진단하고 치료할 때 의사결정을 도와주는 시스템으로 정의한다. Ong 등[43]은 1,359개의 뇌CT, CT혈관조영검사 및 MR혈관조영검사의 판독지를 이용하여 판독지가 뇌경색이 있는지, 중대뇌동맥을 침범하였는지 그리고 급성 병변이 있는지를 예측할 수 있는지 조사하였다. 텍스트를 전처리하는 방법은 "bag-of-words" 모델부터 global vectors for word representation (GloVe) 임베딩 방법을 이용하였고, 자동분류기는 로지스틱회귀분석부터 RNN을 이용한 딥러닝 방법을 모두 적용하였다. 결과적으로 자동분류기의 성능은 GloVe 임베딩을 이용한 순환신경망 알고리즘의 성능이 허혈뇌졸중의 존재 여부, 중대뇌동맥 침범 여부 및 급성기 병변 유무를 예측하는 데 있어서 AUROC 0.93 이상의 성능을 보여주었다. 이와 같이 자연어처리를 이용한 딥러닝 방법을 통하여 급성기 병변의 유무 및 병변의 위치 등을 빠르고 정확하게 분류해낼 수 있는 자연어처리 방법은 비정형텍스트 데이터를 이용한 임상의사결정지원시스템에 많이 사용되고 있다. 이러한 텍스트 판독지를 이용하여 병변의 위치,[38,43-45] 무증상 뇌경색[46]의 여부와 같은 방사선 표지자를 찾은 연구와 더불어, 자연어처리를 임상기록까지 포함하여 분석할 경우 혈관기원의 일과성 뇌허혈에 대한 구분,[47] 뇌졸중의 아형,[48,49] 적절한 정맥혈전용해제 사용을 위한 적응증 및 금기증 확인,[50] 뇌졸중 심각도[51] 및 뇌졸중 3개월 이후의 예후[52] 등의 다양한 표현형을 예측하는 데 있어 빠르고 정확한 임상결정보조정보를 제공할 수 있다. 그러나 현재까지는 텍스트 머신러닝의 성능 또한 개선되어야 할 부분이 있어 텍스트 딥러닝의 활발한 사용을 위해서는 택스트 전처리 방법 및 알고리즘의 고도화가 요구되고 있다.

다른 인공지능 연구분야에 비해 자연어처리 연구에 있어서 더욱 중요한 사항은 개발된 알고리즘에 대한 외부타당도 검사이다. Digital image and communication in medicine 영상이나 심전도의 파형과 같은 시그널 정보와 같이 비교적 정형화된 툴을 기반을 데이터가 수집된 경우에는 고유의 데이터로 개발된 인공지능 파이프라인이 예측이 필요한 다른 데이터(unseen data)에 적용하기 용이하지만, 텍스트라는 것은 판독지나 임상경과기록지가 개인의 성향을 반영한 결과물이며, 구조화되어 있지 않은 자유로운 형태이기 때문에, 개발된 텍스트 기반 알고리즘의 일반화된 성능을 검사하기 위해서는 외부적인 타당성 검사가 반드시 필요하다. 본 뇌졸중 관련 연구들 중에서 이러한 외부타당도가 시행된 연구는 3개의 연구로,[42,43,45] 자연어처리 관련 연구를 해석하고 적용하는 데 있어 이러한 외적 타당도에 대한 증명이 완료된 알고리즘인지를 확인하는 것은 매우 중요하다.

또한 Table에서 시행된 연구들에서는 알고리즘의 성능을 향상시키기 위해 k겹 교차검증(k-fold cross-validation)이 사용되었다. 이 k겹 교차검증은 모델의 학습 및 검증에 사용되는 데이터를 k개로 나누어서 검증과 평가를 시행하는 것을 말한다(Fig. 3). 전체의 데이터는 학습데이터와 검증데이터로 나뉘게 되는데, 이 학습데이터를 k개로 분할한 이후에 k-1개의 겹데이터는 학습에 사용되고

1개의 겹데이터는 알고리즘의 내부타당도를 확인하기 위해 사용되며, 학습이 k번으로 반복이 되면서 학습용데이터로 계발된 알고리즘의 과적합을 최소화할 수 있다.[53]

EMR의 텍스트 머신러닝 분석을 할 때 반드시 고려해야 할 사항은 다루어야 할 텍스트의 종류이다. 대부분의 국내 EMR 기록의 경우 한글과 영문이 혼재되어 사용되고 있다. 철자오류는 많은 데이터를 이용하여 충분히 오류를 극복할 수 있으나 이러한 이중언어문제(bilingualism)는 텍스트 전처리 과정에서 한글 또는 영문만을 따로 전처리할 수밖에 없어 필연적으로 정보의 손실을 가져올 수 있다. 또한 한글의 형태소 분석기의 경우 한글의학용어를 충분히 반영하고 있지 않기 때문에 한영혼재 EMR 기록의 효과적인 전처리는 EMR 자연어처리 분야에서 여전히 해결해야 할 과제 중의 하나이다. 한영혼재 텍스트를 기반으로 한 pretrained word embedding 방법 또는 "네이버지식인"과 같은 한글의학용어의 문장들을 크롤링하여 새로운 의학사전을 만드는 방법 등을 시도해볼 수 있으나 아직까지 개발단계는 미흡한 상황이므로 이러한 한영혼재 EMR 데이터를 다룰 때에는 세심한 주의가 필요하다.

결 론

뇌졸중 대상의 자연어처리 연구가 이제 본격적으로 시작되고 있다. 우리는 자연어처리 방법을 통하여 EMR의 상당부분을 차지하고 있는 비정형텍스트 데이터를 벡터화하고, 컴퓨터가 읽어 들여 자동적으로 중요한 정보를 추출해낼 수 있는 인공지능 알고리즘을 효과적으로 적용할 수 있다. 자연어처리를 통한 기계학습은 인간이 하지 못하는 방대한 정보를 목적에 따라 구분하고, 환자들의 임상경과를 최적화할 수 있는 임상의사결정보조시스템으로 점차 진화할 것이다.

REFERENCES

1. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. *J Int Med* 2013;274:547-560.
2. Jeurgens C. Threats of the data-flood: an accountability perspective in the era of ubiquitous computing. In: Smit F, Glaudemans A, Jonker R. *Archives in liquid times*. s-Gravenhage: Stichting Archiefpublicaties, 2017;196-210.
3. Adnan K, Akbar R, Khor SW, Ali ABA. Role and challenges of unstructured big data in healthcare. In: Sharma N, Chakrabarti A, Balas V. *Data management, analytics and innovation. advances in intelligent systems and computing.* Vol 1042. Singapore: Springer, 2020;301-323.
4. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data processing and text mining technologies on electronic medical records: a review. *J Healthc Eng* 2018;2018:4302425.
5. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. *IEEE Comput Intelli Mag* 2018; 13:55-75.
6. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011;18:544-551.
7. Rajman M, Besançon R. Text mining: natural language techniques and text mining applications. In: Spaccapietra S, Maryanski F. *Data mining and reverse engineering.* Boston: Springer, 1998;50-64.
8. Moreno A, Redondo T. Text analytics: the convergence of big data and artificial intelligence. *IJIMAI* 2016;3:57-64.
9. Nikiforou A, Ponirou P, Diomidous M. Medical data analysis and coding using natural language processing techniques in order to derive structured data information. In: Mantas J, Hasman A. *Informatics, management and technology in healthcare*. Amsterdam: IOS Press, 2013;53-55.
10. Buchlak QD, Esmaili N, Leveque JC, Farrokhi F, Bennett C, Piccardi M, et al. Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review. *Neurosurg Rev* 2020;43:1235-1253.
11. Tan AH. Text mining: the state of the art and the challenges. Proceedings of the PAKDD Workshop on Knowledge Discovery from Advanced Databases; 1999; Beijing, China. p. 65-70.
12. Brants T. Natural language processing in information retrieval. Proceedings of CLIN 2003; 2003; Antwerp, Belgium. Netherland: Schloss Dagstuhl - Leibniz Center for Informatics; 2004. p. 1-13.
13. Vijayarani S, Ilamathi MJ, Nithya M. Preprocessing techniques for text mining-an overview. *Int J Comput Sci Commun Netw* 2015;5:7-16.
14. HaCohen-Kerner Y, Miller D, Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation. *PLoS One* 2020;15:e0232525.
15. Willett P. The porter stemming algorithm: then and now. *Program* 2006;40:219-223.
16. Ferilli S, Esposito F, Grieco D. Automatic learning of linguistic resources for stopword removal and stemming from text. *Procedia Comput Sci* 2014;38:116-123.
17. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing rule-based and deep learning models for patient phenotyping. [online] [cited 2021 Mar 24]. Available from: URL: https://arxiv.org/abs/1703.08705.
18. Eftimov T, Koroušić Seljak B, Korošec P. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS One* 2017;12:e0179488.
19. Aronson AR. Metamap: mapping text to the umls metathesaurus. [online] [cited 2021 Mar 24]. Available from: URL:https://lhncbc.nlm.nih.gov/ii/information/Papers/metamap06.pdf.
20. Chiang JH, Lin JW, Yang CW. Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE). *J Am Med Inform Assoc* 2010;17:245-252.
21. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507-513.
22. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*; 2001; Washington, DC. Oxford: Oxford University Press;

2004. p. 17.
23. Jing L, Ng MK, Huang JZ. Knowledge-based vector space model for text clustering. *Know Inform Syst* 2010;25:35-55.
24. Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity. *Aaai* 2006;6:775-780.
25. Zhang Y, Jin R, Zhou ZH. Understanding bag-of-words model: a statistical framework. *Int J Mach Learn Cybernet* 2010;1:43-52.
26. Voorhees EM. Natural language processing and information retrieval. In: Pazienza MT. *Information extraction.* Heidelberg: Springer, 1999; 32-48.
27. Cavnar WB, Trenkle JM. N-gram-based text categorization. Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval; 1994. Las Vegas, NV: University of Nevada; 1994, p. 161-175.
28. Goldberg Y, Levy O. word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. [online] [cited 2021 Mar 24]. Available from: URL:https://arxiv.org/abs/1402.3722.
29. Bianchi B, Monzón GB, Ferrer L, Slezak DF, Shalom DE, Kamienkowski JE. Human and computer estimations of predictability of words in written language. *Sci Rep* 2020;10:1-11.
30. Liu Q, Huang HY, Gao Y, Wei X, Tian Y, Liu L. Task-oriented word embedding for text classification. Proceedings of the 27th international conference on computational linguistics; 2018; Santa Fe, NM. Cambridge: Massachusetts Institute of Technology Press; 2018. p. 2023-2032.
31. Cer D, Yang Y, Kong SY, Hua N, Limtiaco N, John RS, et al. Universal sentence encoder. [online] [cited 2021 Mar 24]. Available from: URL:https://arxiv.org/abs/1803.11175.
32. Kang N, Singh B, Afzal Z, van Mulligen EM, Kors JA. Using rule-based natural language processing to improve disease normalization in biomedical text. *J Am Med Inform Assoc* 2013;20:876-881.
33. Razno M. Machine learning text classification model with NLP approach. *Comput Ling Intellig Syst* 2019;2:71-73.
34. Kanakaraj M, Guddeti RMR. Performance analysis of ensemble methods on twitter sentiment analysis using NLP techniques. Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015); 2015; Anaheim, CA. New York, NY: Institute of Electrical and Electronics Engineers; 2015. p. 169-170.
35. Li J, Chen X, Hovy E, Jurafsky D. Visualizing and understanding neural models in nlp. [online] [cited 2021 Mar 23]. Available from: URL: https://arxiv.org/abs/1506.01066.
36. Li H. Deep learning for natural language processing: advantages and challenges. *Nat Sci Rev* 2017;5:24-26.
37. Vosoughi S, Vijayaraghavan P, Roy D. Tweet2vec: learning tweet embeddings using character-level cnn-lstm encoder-decoder. Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval; 2016; Pisa, Italy. New York, NY: Association for Computing Machinery; 2016. p. 1041-1044.
38. Elkins JS, Friedman C, Boden-Albala B, Sacco RL, Hripcsak G. Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. *Comput Biomed Res* 2000;33:1-10.
39. Pons E, Braun LM, Hunink MM, Kors JA. Natural language processing in radiology: a systematic review. *Radiology* 2016;279:329-343.
40. Mowery DL, Chapman BE, Conway M, South BR, Madden E, Keyhani S, et al. Extracting a stroke phenotype risk factor from Veteran Health Administration clinical reports: an information content analysis. *J Biomed Seman* 2016;7:1-12.
41. Kim C, Zhu V, Obeid J, Lenert L. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. *PLoS One* 2019;14:e0212778.
42. Li MD, Lang M, Deng F, Chang K, Buch K, Rincon S, et al. Analysis of stroke detection during the COVID-19 pandemic using natural language processing of radiology reports. *AJNR Am J Neuroradiol* 2021; 42:429-434.
43. Ong CJ, Orfanoudaki A, Zhang R, Caprasse FP, Hutch M, Ma L, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. *PLoS One* 2020;15:e0234908.
44. Alex B, Grover C, Tobin R, Sudlow C, Mair G, Whiteley W. Text mining brain imaging reports. *J Biomed Seman* 2019;10:1-11.
45. Wheater E, Mair G, Sudlow C, Alex B, Grover C, Whiteley W. A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records. *BMC Med Inform Decis Mak* 2019;19:1-11.
46. Fu S, Leung LY, Wang Y, Raulli AO, Kallmes DF, Kinsman KA, et al. Natural language processing for the identification of silent brain infarcts from neuroimaging reports. *JMIR Med Inform* 2019;7:e12109.
47. Bacchi S, Oakden-Rayner L, Zerner T, Kleinig T, Patel S, Jannes J. Deep learning natural language processing successfully predicts the cerebrovascular cause of transient ischemic attack-like presentations. *Stroke* 2019;50:758-760.
48. Garg R, Oh E, Naidech A, Kording K, Prabhakaran S. Automating ischemic stroke subtype classification using machine learning and natural language processing. *J Stroke Cerebrovasc Dis* 2019;28:2045-2051.
49. Sung SF, Lin CY, Hu YH. EMR-based phenotyping of ischemic stroke using supervised machine learning and text mining techniques. *IEEE J Biomed Health Inform* 2020;24:2922-2931.
50. Sung SF, Chen KC, Wu DP, Hung LC, Su YH, Hu YH. Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: a feasibility study. *Int J Med Inform* 2018;112:149-157.
51. Kogan E, Twyman K, Heap J, Milentijevic D, Lin JH, Alberts M. Assessing stroke severity using electronic health record data: a machine learning approach. *BMC Med Inform Decis Mak* 2020;20:8.
52. Heo TS, Kim YS, Choi JM, Jeong YS, Seo SY, Lee JH, et al. Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI. *J Pers Med* 2020;10:286.
53. Cawley GC, Talbot NL. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 2010;11:2079-2107.
54. Guan W, Ko D, Khurshid S, Trisini Lipsanopoulos AT, Ashburner JM, Harrington LX, et al. Automated electronic phenotyping of cardioembolic stroke. *Stroke* 2021;52:181-189.