자연어분석에 대한 경험

김 철 호[ab], 배 종 석[bc]
춘천성심병원 신경과[a], 한림대학교 신경과학교실[b], 강동성심병원 신경과[c]

Experience of Natural Language Processing

Chulho Kim[ab], MD, PhD, Jong Seok Bae[bc], MD, PhD
[a]Department of Neurology, Chuncheon Sacred Heart Hospital, Chuncheon, Korea, [b]Department of Neurology, Hallym University College of Medicine, Chuncheon, Korea, [c]Department of Neurology, Kangdong Sacred Heart Hospital, Seoul, Korea

Natural language processing (NLP) is a subfield of computer science, information engineering, and artificial intelligence (AI) concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.[1] Recently, the application of big data analytics in healthcare has a lot of positive and also life-saving outcomes using unstructured data, such as imaging or text data. Herein, we will briefly investigate what NLP is and how we can apply it to the medical field we have.

Introduction

Natural language processing (NLP) is a subfield of computer science, information engineering, and artificial intelligence (AI) concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. Recently, the application of big data analytics in healthcare has a lot of positive and also life-saving outcomes using unstructured data, such as imaging or text data. Herein, we will briefly investigate what NLP is and how we can apply it to the medical field we have.

Jong Seok Bae
Department of Neurology, Kangdong Sacred Heart Hospital, Seoul, Korea
Seongan-ro, Gangdong-gu, Seoul, 05355, Republic of Korea
Tel: +82-2-2224-2206
Fax: +82-2-478-6330
E-mail: jsb_res@hotmail.co.kr

1. How can we make text data available for analysis?

Text can be converted into data through various methods. Here we describe two methods, tokenization and part-of-speech (POS) speech tagging, which are the most frequently used methods for the text vectorization. First, "tokens" are usually individual words (at least in languages like English) and "tokenization" is taking a text or set of text and breaking it up into its individual words (Fig.1).[2] The whole set of tokens is defined as corpus, and the corpus defines the document to be analyzed as not being annotated.[3] These tokenized words (tokens) can then be used to create vectors in the same way as bag-of-words (BOW) models or skip-grams.[4] The BOW model treats individual words as a vector without regard to the order, while skip-gram is a method of calculating the probability

that a word is located according to a sequence in a vector. POS tagging refers to the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context—i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc (Fig. 2).[5]

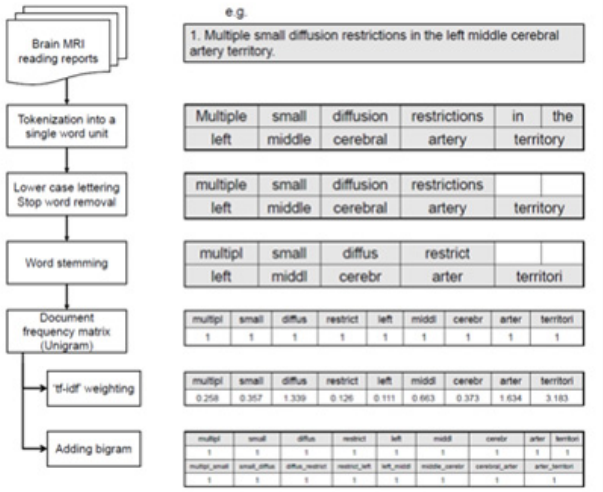
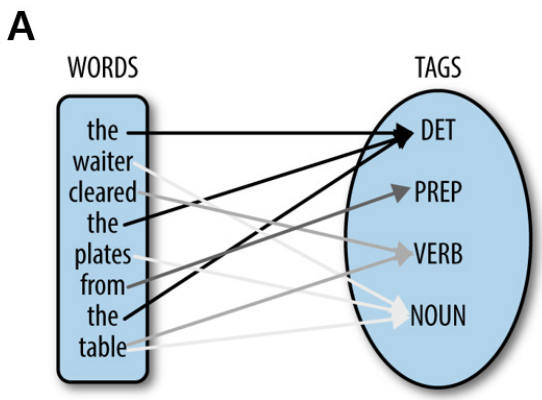


Fig 1. Example of tokenization of radiology report documents.
Full text brain MRI reading sentences were initially parsed into "tokens," with numbers, punctuations, symbols and hyphens in the original text data removed. Then, other unnecessary tokens were removed using lowercase lettering, stop word removal, and word stemming to normalize those data. Finally, we obtained the document-feature matrix (dfm), which is a vector representation of tokens that are truncated from the whole text. Term frequency-inverse document frequency (tf-idf) weighting and adding bigram can be applied to these vectors.

2. Annotation method according to the task

Any metadata tag used to mark up elements of the dataset is called an annotation over the input. In case of supervised machine learning (ML), golden standard labeling (or annotation) of the document could be differ according to the ML tasks. Labeling can be continuous variable if we use regression ML task, or di-/poly-chotomous variables in classification task. In automated annotation of NLP field, we can use word2vec or text2vec materials. The word2vec software of Tomas Mikolov and colleagues has gained a lot of attention recently, and provides state-of-the-art word embeddings in NLP tasks.[6] Word2vec is a group of related models that are used to produce word embeddings. Word2vec takes as its input a large corpus of text and produces a vector space, typically of several hundred dimensions, with each unique word in the corpus being assigned a corresponding vector in



B

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

Tag	Description
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP$	Possessive whpronoun
WRB	Whadverb

Fig 2. Part-of-speech tagging
Figure A illustrate an example of part-of-speech tagging methods, and B shows the standard OpenNLP taggers.

the space. We can use these predetermined word embeddings such as word2vec or task-specific word embedding which can be obtained from the corpus directly.[7]

3. Supervised machine learning

We can predict our ML algorithm's performance using F1-score, area under the ROC (AUROC), precision, recall and accuracy (Fig. 3). In binary classification, precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score. We used superficial or deep learning as we want to perform. At first, we divide the whole document dataset into training and test dataset with the ratio of 8:2. There is no golden standard rule in dividing training and test dataset, but division ratio can be adjusted according to the trade-off between false positives and false negatives in the training dataset.[8] Second, we can obtain the probabilities of each training dataset. Finally, we calculated performance score of the test dataset in predicting true positive, false positive, false negative, and true negative cases. In case of superficial ML, such as decision tree, random forest, and extreme gradient boosting, we can intuitively understand what tokens have influenced the performance of automatic classification machine learning (Fig. 4)

$$\text{F1-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \text{Positive Predictive Value} = \text{TP}/(\text{TP}+\text{FP})$$

$$\text{Recall} = \text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN})$$

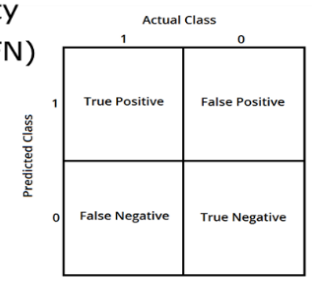


Fig 3. Several types of performance measure used in ML prediction
TP, true positive; FP, false positive; FN, false negative; TN, true negative.

4. Future perspectives.

NLP have been used in a variety of computational fields such as machine translation, information retrieval and extraction, and robot journalism. Especially, automated data tracking and analysis software in radiology fields has

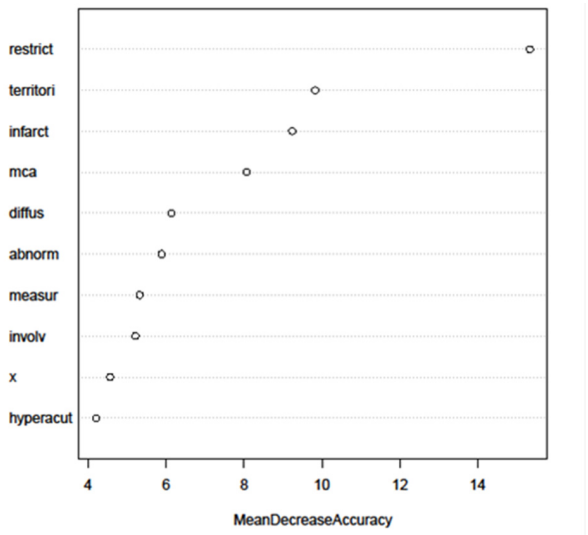


Fig 4. An example of variable importance plot in random forest ML classification
Y axis represents the stemmed word unigram vectors in the random forest prediction algorithm. X-axis represents the mean decreased accuracy of the word vectors. For example, the stemmed word 'restrict' is the most valuable word vectors to distinguish whether the brain MRI report refer to acute ischemic stroke or not.

been increased, the implementation of NLP tools in radiology field NLP continues to expand to the other fields. In addition, the accuracy of NLP prediction is getting better when we use deep learning algorithm. Therefore, it is expected that the decision support system using this automatic prediction algorithm will be developed more.

References

1. Manning CD, Manning CD, Schütze H. Foundations of statistical natural language processing:MIT press 1999.
2. Buyko E, Wermter J, Poprat M, Hahn U. Automatically adapting an NLP core engine to the biology domain. Proceedings of the ISMB 2006:65-8.
3. Carroll J, Minnen G, Briscoe T. Corpus annotation for parser evaluation. arXiv preprint cs/9907013 1999
4. Liu P, Qiu X, Huang X. Learning Context-Sensitive Word Embeddings with Neural Tensor Skip-Gram Model. IJCAI 2015:1284-90.
5. Rajput A. Natural Language Processing, Sentiment Analysis and Clinical Analytics. arXiv preprint arXiv:1902.00679 2019
6. Wu L, Yen IE, Xu K, Xu F, Balakrishnan A, Chen P-Y, et al. Word Mover's Embedding: From Word2Vec to Document Embedding. arXiv preprint arXiv:1811.01713 2018
7. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. arXiv preprint arXiv:1802.05365 2018
8. Raschka S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. 2018