Analyzing Medical Records Using Text Mining

08.31.2019
Min Song, Ph. D.
Department of Library and Information Science
Department of Digital Analytics
Yonsei University

CONTENT

1 Introduction

01 Background

- Text is a very important form of data in the biomedical field.

- For example, patient records contain large amounts of text which has been entered in a non-standardized format, consequently posing a lot of challenges to processing of such data (Holzinger et al., 2014).

- However, the steadily increasing volumes of unstructured information need machine learning approaches for data mining, i.e. text mining (Holzinger et al., 2014).

02 Necessity of clinical record analysis

- Mining hospital data holds the potential for new discoveries as well as for enabling improved efficiency and communication within hospital systems (Kocbek et al., 2016).

- Much valuable information in hospital records is represented in free text format, e.g., radiology and pathology reports, requiring the application of text mining and Natural Language Processing (NLP) techniques (Kocbek et al., 2016).

- In this study, we applied text mining techniques to X-ray image reading to implement a system to automatically determine whether a particular medical record is to be labeled as pneumonia or not.

03 Research Objectives

- First main goal is **to find optimal text representation for automatic clinical doctor's note classification**.
  - Introducing the 'TF-IDF ratio' to transform the original text data
  - Applying N-gram technique to enhance matching probabilities between sample text and all document set
  - Using Word2Vec model to expand our refined data by TF-IDF ratio and N-gram

- Second main goal is **to develop the deep learning algorithm that is most suitable for the optimized text representation**.
  - Proposing a model to vectorize each modified medical record using Doc2Vec and then applying CNN

2 Related Work

01 Text Mining Based Medical Analysis

- Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). *Data processing and text mining technologies on electronic medical records: a review*. Journal of healthcare engineering, 2018.
  - EMR(Electronic Medical Records) is unstructured data, it requires data cleansing, data integration, data transformation, data reduction. It is necessary to preprocess the source data in order to improve data quality and improve the data mining results.

- Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). *What can natural language processing do for clinical decision support?*. Journal of biomedical informatics, 42(5), 760-772.
  - CDS(Clinical Decision Support) aims to provide easily accessible health-related information. This study primarily focuses on development of fundamental NLP(Natural Language Processing) methods and advances in the NLP systems for CDS.

02 Deep learning based medical record analysis

- Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). *Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis*. IEEE journal of biomedical and health informatics, 22(5), 1589-1604.
  - EHR(Electronic Health Record) data, where we find a variety of deep learning techniques and frameworks being applied to several types of clinical applications including information extraction, representation learning, outcome prediction, phenotyping, and de-identification.

- Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., ... & Aerts, H. J. (2019). *Deep learning predicts lung cancer treatment response from serial medical imaging.* Clinical Cancer Research, 25(11), 3266-3275.
  - 179 patients' CT images are predicted by using deep learning techniques; CNN (Convolutional Neural Networks) and RNN (Recurrent Neural Networks).

3 Method

01 Overview of Method

- It was pivotal **to make a training dataset that properly represents clinical doctor's notes** and compare the performance between raw sentence data and represented data.

- And **we developed and proposed a new deep learning algorithm to automatically classify reconstructed text** through the above operations.

01 Overview of Method

Data Collection

- X-ray image reading : 1,275
  → classified with 3 criteria
  (Not Pneumonia / Pneumonia / Judgement Hold)

Pre-processing

- Stopword Removal (Eng, Kor)
- Special character processing
- Lowercase

Data Transformation

- Raw data
- TF-IDF ratio
- N-gram
- Word Embedding (word2vec)

PKDE4J

- Named Entity Recognition
- Relation Extraction

Classification

- Vectorization: Doc2Vec
  (500 dimensions, 0.025 learning rate, 1000 batch size, and 20 epochs)
- Combined with CNN
  (Doc2vec + CNN)

Evaluation

- Raw Data
- Raw Data + Word2Vec
- Raw Data + N-gram
- Raw Data + N-gram + Word2Vec
- TF-IDF ratio
- TF-IDF ratio + Word2Vec
- TF-IDF ratio + N-gram
- TF-IDF ratio + N-gram + Word2Vec

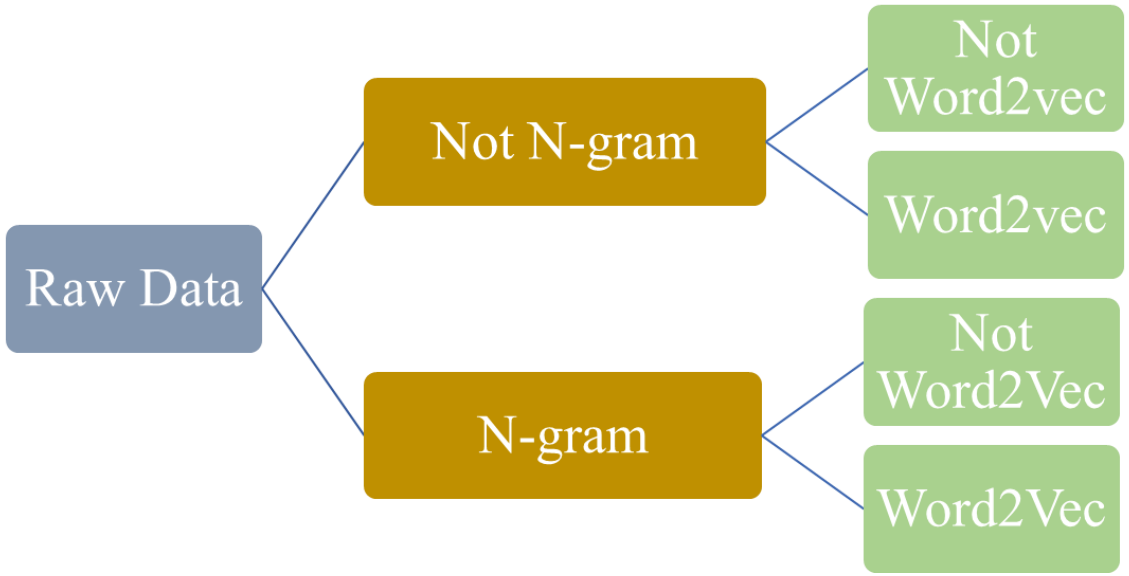
'Doc2Vec + CNN' algorithm was commonly applied to all cases.

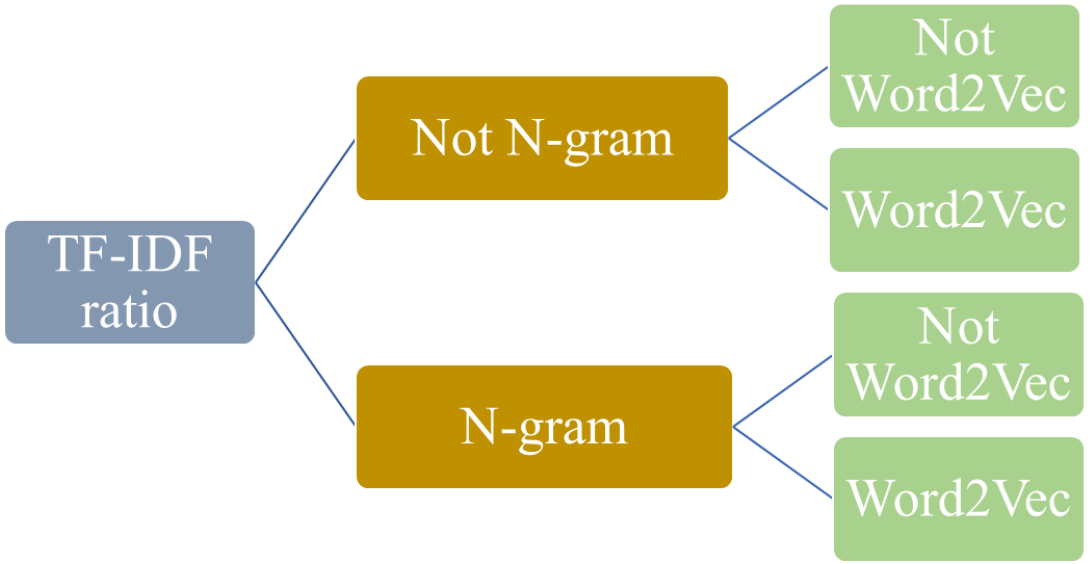
Web Application

- Annotated entity and relation
- Chart

01 Overview of Method

- Case of evaluation





‘Doc2Vec + CNN’ algorithm was commonly applied to all cases.

02 Data collection

- A total of 1,275 X-ray image reading data were provided from Asan Hospital.

Pneumonia judgment result	Number of data
Not Pneumonia	673
Pneumonia	371
Judgement Hold	231
Total	**1,275**

- Of these, 80% of the data was used to create the model as training data, and the remaining 20% was used to evaluate the performance of the trained model as test data.

03 Sample of training data

In the column on the capture below,

- description2: X-ray image reading
- pneumonia: Pneumonia or not (3 categories)
  - 0: Not Pneumonia / 1: Pneumonia / 2: Judgement Hold

1	Unnamed:	idx	adm_day	discharge_	xray_day	xray_t	orderday	adm2_day	code	code_name	descrip	result	description2	pneumonia
4754	3875	12333	1168891048	20110609	20110615	20110609	1635	20110609	20110609	RD3000	CR, Chest P	CR, Chest PA,	**NO change of nod**	**0**
4755	12008	46538	1173512555	20111021	20111102	20111029	658	20111029	20111021	RD3001	CR, Chest P	CR, Chest PA	**No interval change**	**0**
4756	10724	40538	1172884804	20170926	20171002	20170930	1338	20170930	20170926	RD3001	CR, Chest P	CR, Chest PA	**Postop left hydrop**	**0**
4757	12120	46934	1173579632	20161228	20170205	20170113	2	20170113	20161228	RD3001	CR, Chest P	CR, Chest PA	**1. No change of fil**	**1**
4758	1461	3844	1167555697	20120111	20120118	20120113	1628	20120113	20120111	RD3001	CR, Chest P	CR, Chest PA	**S/P left lower lobe**	**0**
4759	13155	52349	1174141075	20110127	20110331	20110320	1312	20110320	20110127	RD3001	CR, Chest P	CR, Chest PA	**No change of puln**	**0**
4760	11743	45601	1173385990	20131014	20131028	20131021	550	20131021	20131014	RD3001	CR, Chest P	CR, Chest PA	**s/p wedge resectio**	**0**
4761	16465	64712	1175887109	20140723	20140731	20140723	1705	20140723	20140723	RD3000	CR, Chest P	CR, Chest PA	**s/p lobectomy, RU**	**0**
4762	25031	110749	1180426276	20100331	20100407	20100402	1309	20100402	20100331	RD3001	CR, Chest P	CR, Chest PA	**S/P RUL lobectomy**	**0**
4763	25222	111429	1180533944	20170221	20170226	20170224	530	20170224	20170221	RD3001	CR, Chest P	CR, Chest PA	**S/P RLLobectomy (**	**0**
4764	7261	26152	1170860603	20130903	20130928	20130914	1008	20130914	20130903	RD3001	CR, Chest P	CR, Chest PA	**Minimal subsegme**	**0**
4765	23996	104729	1179943322	20111009	20111105	20111019	554	20111019	20111009	RD3001	CR, Chest P	CR, Chest PA	**S/P RLL lobectomy**	**0**
4766	5471	17746	1169749655	20131124	20140108	20131127	2048	20131127	20131124	RD3001	CR, Chest P	CR, Chest PA	**No active lesion in**	**0**
4767	24224	106013	1180049108	20090531	20090809	20090805	1526	20090805	20090531	RD3001	CR, Chest P	CR, Chest PA	**Both pleural effusi**	**2**

03 Sample of training data

- Red letters indicate whether patients have pneumonia

Sample X-ray image reading	Pneumonia or not
S/P right lower lobectomy on 2013-9-28 for lung cancer. Right chst tube insertion state. Small amount of subcutaneous emphysema along the right chest wall. ***Increased extent of patchy opacity in BLLZ***, since 2013-9-28. -->R/O aspiration pneumonia. -->Clinical correlation.	Pneumonia
Tracheostomy state.Pig-tail insertion state in right hemithorax. ***Diffuse haziness in both lungs***. Increase of interstitial marking. ***Consolidation and peribronchial infiltration mainly in left central lung. Multifocal irregualar peribronchial infiltrations in both lungs***.--> Pulmonary edema with pleural effusion. R/O combined pneumonia.--> Wax and wine on serier f/u images.	Possible Pneumonia
History:1. Biopsy proven SqCC.2. Compared with 2016-05-05 CT. Findings and Conclusions:1. A bout 1.7 cm lobulated irregular nodule in RLL, posterobasal segmental bronchus with peripheral patchy collapse and consolidation. - sl. decreased extent of peripheral peribronchial consolidation. - newly developed adjacent pleural thickening, r/o reactive change. - decreased right pleural effusion. - enlarged right interlobar LNs.---> biopsy proven lung cancer with ***postobsructive pneumonia*** (T1a N1 Mx).2. Emphysema, both lungs.3. Atherosclerotic plaque and calcification in aorta.4. No abnormal finding in bony thorax.	Not Pneumonia

Module Explanation

<Module 1> Pre-processing,
Data Transformation

01 Pre-processing

- Special Character Processing
- Lowercase
- Stop-words Removal (Eng, Kor)
  - When we checked the actual written findings, no meaningful Korean term was found.

02 Data Transformation

1) Text Representation

- To restructure the literature into words that better reveal the characteristics of the category
- Introducing a new type of ‘TF-IDF ratio’

* TF-IDF (Term Frequency – Inverse Document Frequency)

: **Statistical weights that indicate how important a word is in a document** when you have a set of documents.

TF (Term Frequency)

: The number of occurrences of the word in the document.

IDF (Inverse Document Frequency)

: An indicator of the scarcity of words in the entire literature group.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Therefore, **words with a higher TF-IDF value are keywords that distinguish one document from another.**

02 Data Transformation

1) Text Representation

- Based on this, we introduced a new concept.

(1) Finding TF-IDF values for all words in over 10,000 documents

→ Named '**Global TF-IDF**'

(2) Calculating TF-IDF value of three categories from 252 samples

→ Name it '**Local TF-IDF**'

02 Data Transformation

1) Text Representation

- We need to find the ratio so you can calculate exactly **how important a word is in this particular category**.

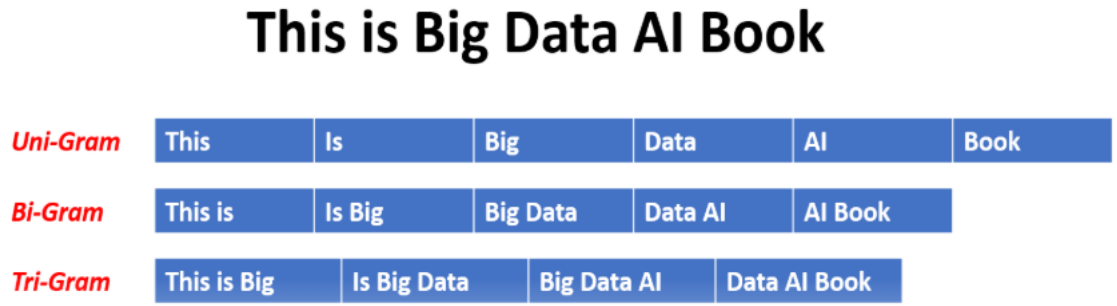
- $\frac{\textbf{Local TF}-\textbf{IDF}}{\textbf{Global TF}-\textbf{IDF}}$ **:** The ratio of **the relative importance of each word** is derived

- **By leaving only words that have a certain level of TF-IDF ratio**, the raw text is refined.

02 Data Transformation

2) Applying N-gram technique

- In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sample of text or speech.
- N-gram examples from various disciplines



Field	Unit	Sample sequence	1-gram sequence	2-gram sequence	3-gram sequence
Vernacular name			Unigram	Bigram	Trigram
Protein sequencing	amino acid	… Cys-Gly-Leu-Ser-Trp …	…, Cys, Gly, Leu, Ser, Trp, …	…, Cys-Gly, Gly-Leu, Leu-Ser, Ser-Trp, …	…, Cys-Gly-Leu, Gly-Leu-Ser, Leu-Ser-Trp, …
DNA sequencing	base pair	…AGCTTCGA…	…, A, G, C, T, T, C, G, A, …	…, AG, GC, CT, TT, TC, CG, GA, …	…, AGC, GCT, CTT, TTC, TCG, CGA, …

02 Data Transformation

2) Applying N-gram technique

- **The N-gram index method has the advantage that any word in the original document is searched for, and no search misses**.

- For example, a search word containing a stop word can't be searched by the word index method but can be searched by the N-gram index method.

- This allows us **to make up for the lack of sample data and lay the foundation for more precise classification**.

02 Data Transformation

3) Word embedding-based data expansion

- Mapping training data to PubMed using **Word2Vec** algorithm
  - Word2vec was created by a team of researchers led by Tomas Mikolov at Google and patented. (Tomas Mikolov et al., 2013)
  - Word2vec is a group of related models that are used to produce word embeddings.
  - Word vectors are positioned in the vector space such that words that share common contexts in the corpus are located in close proximity to one another in the space.

- While the two techniques described in the previous section were intended to highlight the characteristics of each category, or to increase the precision of the search, **this method is a countermeasure to dramatically increase the amount of excessive training data itself**.

Module Explanation

<Module 2> Vectorization,
Deep learning based automatic classification

01 Vectorization

- **Doc2vec** model (Le & Mikolov, 2014; Nebot et al., 2018)
  - **Converting to a vector value for a large block of text, such as the body of a news article.**
  - Think of it as **an extended version of Word2vec**.
  - In the Word2Vec model, there were two important algorithms (CBOW, Skip-gram). Similarly, there are Distributed Memory (DM) and Distributed Bag of Words (DBOW) in the Doc2Vec model.
  - The input to Doc2Vec is an iterator of Labeled Sentence objects.
  - Each object represents a sentence and consists of a list of words and a list of labels.

01 Vectorization

- **We transformed optimized medical records through ‘Module 1’ into sentences through embedding by Doc2Vec to represent sentences with linguistic as well as semantic properties and rules.**

- The assumption of this algorithm is that Doc2Vec makes similar document vector representations within the same category.

- We made the input for the Doc2vec model as a vector format based on the original form of sentences with dimensionality of 500, a learning rate of 0.025, a batch size of 1000, and 20 epochs.

02 Deep learning based automatic classification

- **Convolutional neural network (CNN)** (LeCun & Bengio, 1995).
  - The CNN consists of a multiple convolutional layer, a pooling layer, and a fully connected layer.
  - It is best suited to train two-dimensional data, and it is usually applied for image recognition (LeCun & Bengio, 1995).

02 Deep learning based automatic classification

Doc2Vec + CNN

- In this model, we applied the Doc2vec model as an input feature for CNN to overcome the inefficient zero-padding problem when using CNN with raw text. The CNN sets the sentence length as equal, and this can cause the zero-padding problem.

- **The input was applied to the CNN combined Doc2vec model**.
- In the model building process, we set the layer size to 500 with a fully connected multi-dense layer and one softmax output layer.

- **It can learn contextual information in a document**.

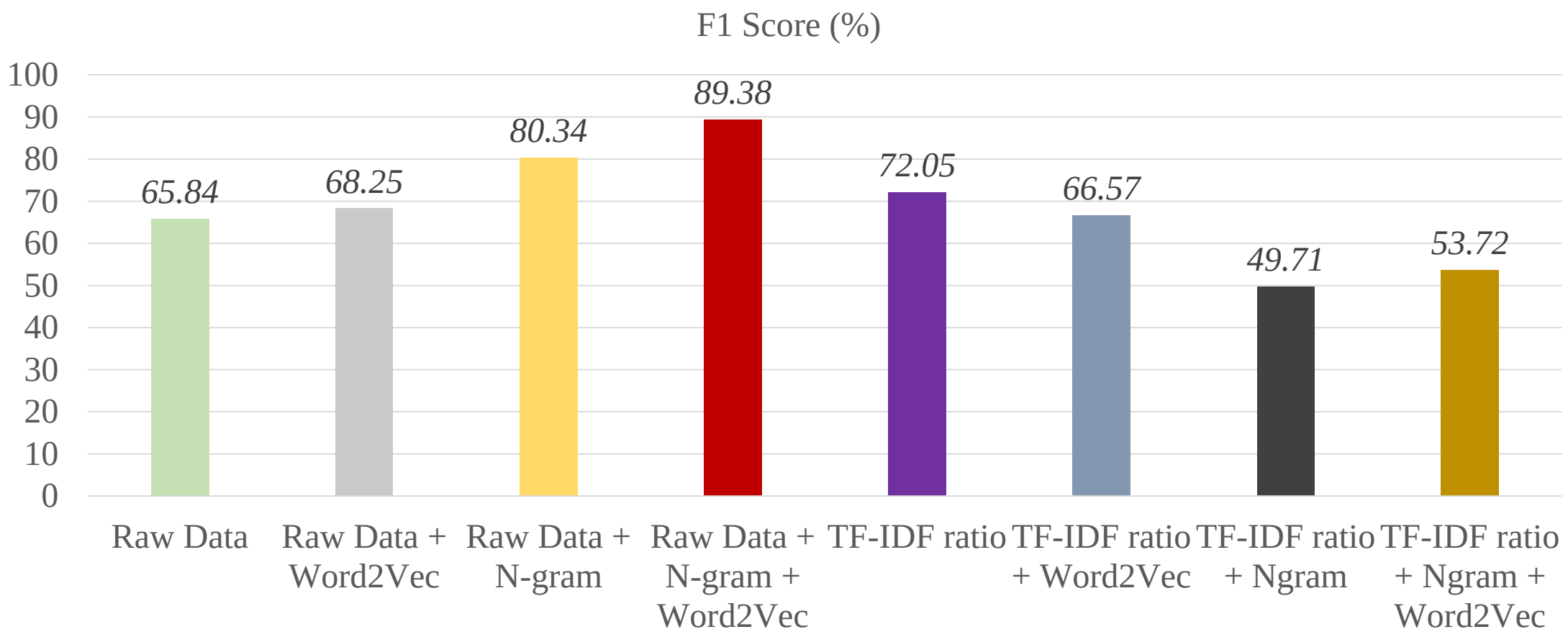
4 Result

O1 Performance experiment

To evaluate performance of our methodology, experiments were conducted considering **8 cases**. As mentioned earlier, **'Doc2Vec + CNN' algorithm was commonly applied** to all cases.

Shape of represented data	Applied classification algorithm
1) Raw Data	**Doc2Vec + CNN**
2) Raw Data + Word2Vec	
3) Raw Data + N-gram	
4) Raw Data + N-gram + Word2Vec	
5) TF-IDF ratio	
6) TF-IDF ratio + Word2Vec	
7) TF-IDF ratio + N-gram	
8) TF-IDF ratio + N-gram + Word2Vec	

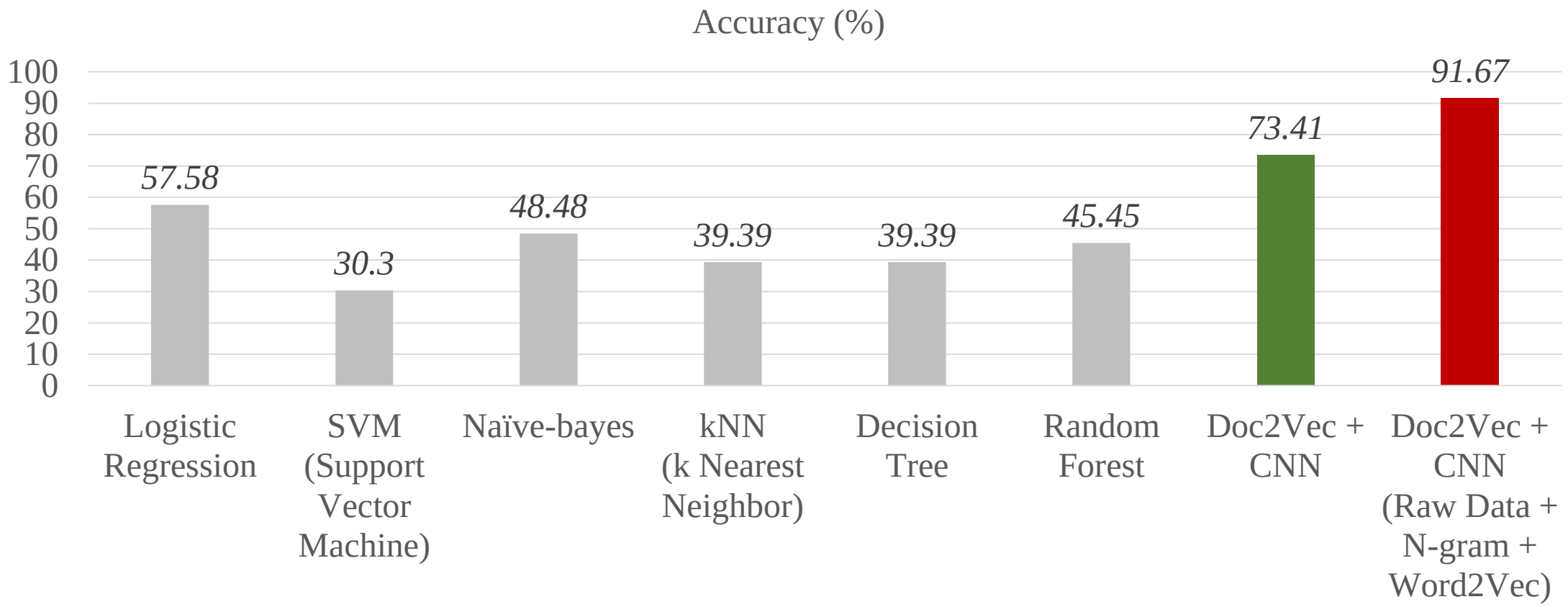
02 Performance evaluation



- **'Doc2Vec + CNN' algorithm was commonly applied** to all cases.

02 Performance evaluation

- Data type: Raw Data (No data transformation)



- Compared to many other machine learning algorithms, our '**Doc2Vec + CNN**' scores significantly higher accuracy even under the common condition of raw data.

03 Interpretation of performance among datasets

- **Throughout experiments, the case of Raw Data expanded by Word2Vec with n-gram (1, 2, 3 gram) transformation shows the highest performance (F-score: 0.894).**

- Without considering the number of tokens in the document, Doc2vec represents a document itself as a vector and transforms text as well as category label simultaneously.

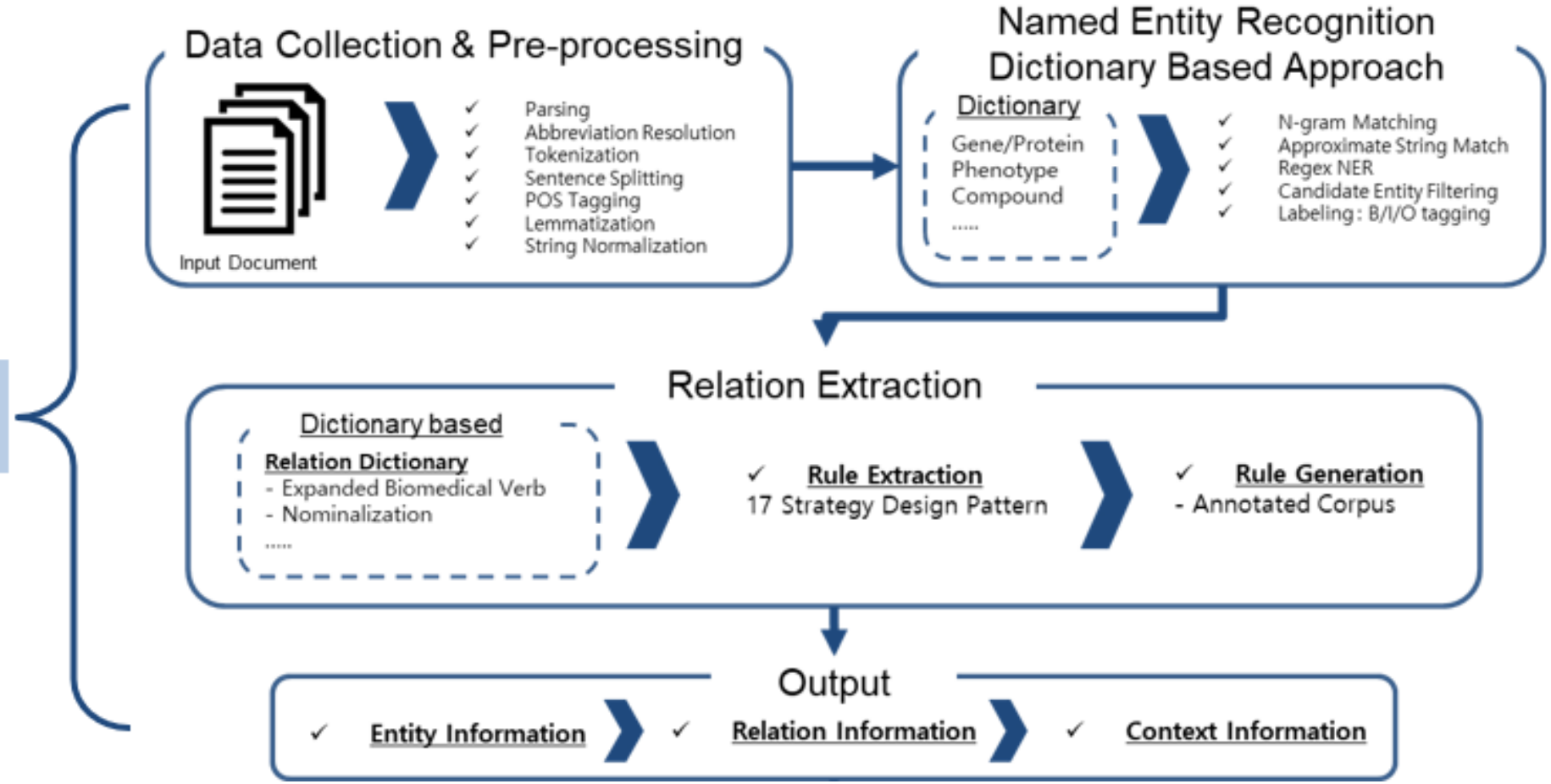
- Therefore, it can learn contextual information in a document.
- The original dataset conveys rich contextual terms, and this is why the CNN combined with Doc2Vec can learn better using this dataset.

5 *Exploration Engine*

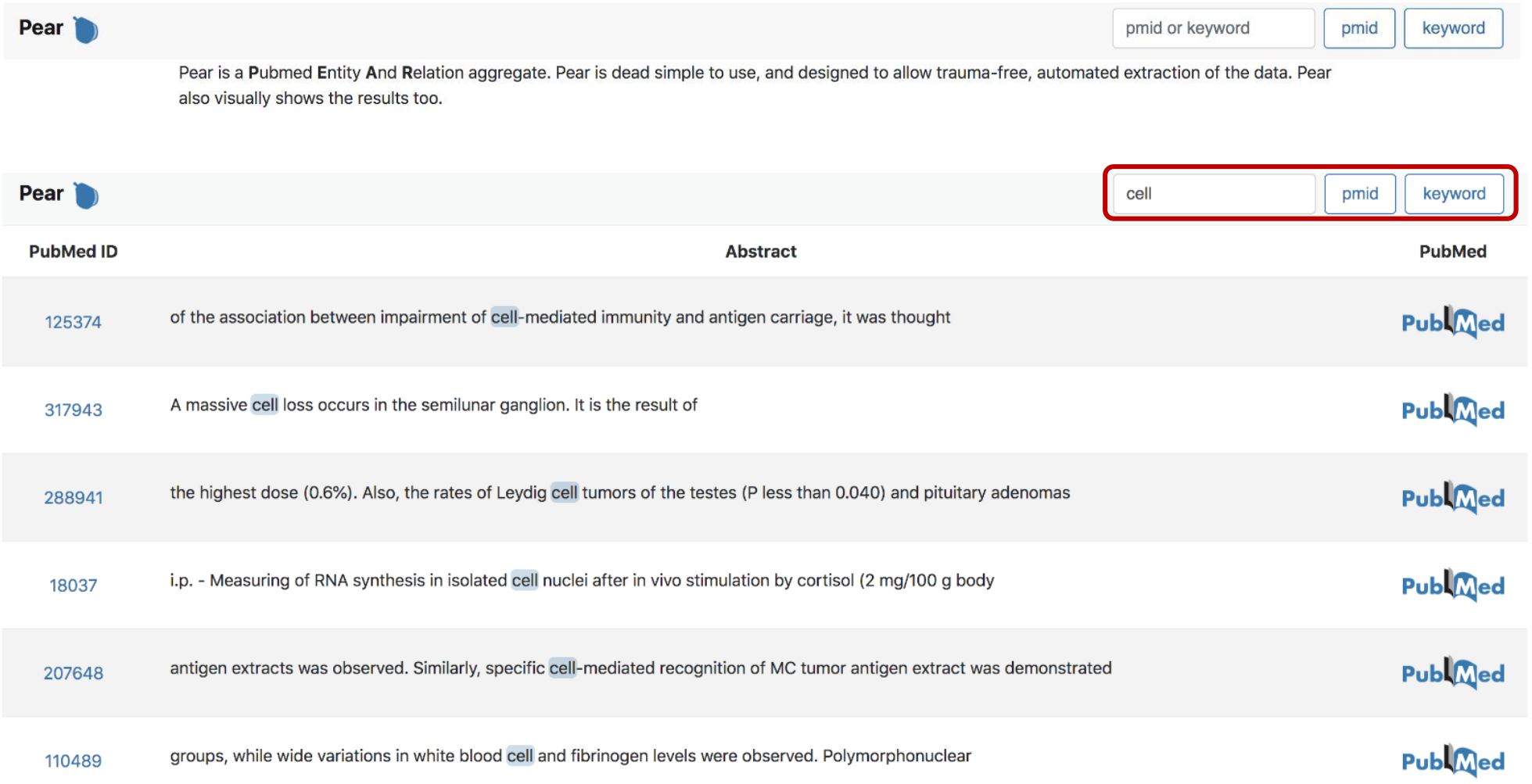
01 Entity and Relation extraction by PKDE4J

Workflow

PKDE4J 2.0
- Performance Improvement



02 Web Application

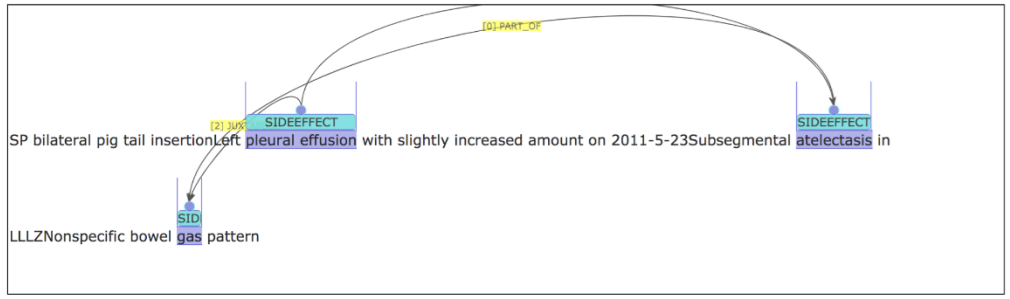


02 Web Application

Pear

pmid or keyword | pmid | keyword

Vizualization



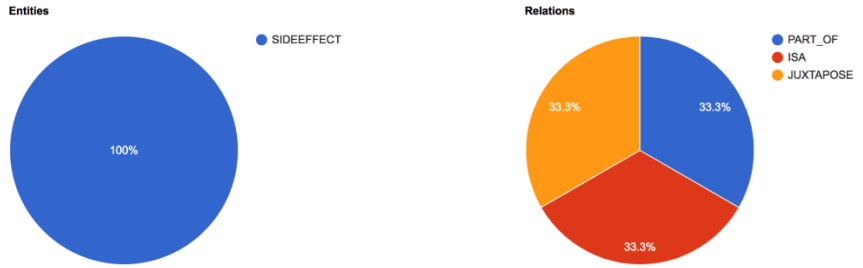
Entities

sentence id	entity id	entity text	entity type	entity span
1	0	atelectasis	SIDEEFFECT	13 13
1	1	gas	SIDEEFFECT	17 17
1	2	pleural effusion	SIDEEFFECT	5 6

Relations

sentence id	relation id	relation type	subject	predicate
1	0	PART_OF	atelectasis	gas
1	1	ISA	pleural effusion	atelectasis
1	2	JUXTAPOSE	pleural effusion	gas

Charts



6 Conclusion

01 The Need for Text Mining in the Medical Field

TEXT MINING PATIENT RECORDS:

EXTREMELY COMPLICATED BUT INCREDIBLY REWARDING

- Text mining of medical records provides a good example of how text mining can make a difference in the real world.
- It also provides a good example of some of the problems encountered, and of the solutions that we may need to deploy.

- As we have shown in this study, **highly accurate classifiers and exploratory analytics tools are essential to obtain meaningful insights from doctors' medical records through text mining techniques.**

02 Importance of proper text representation

- Although improving the performance of a given task by applying a specific algorithm is important of course, but **it is equally important to find a text representation optimized for that algorithm**.

- Therefore, **efforts should always be made to understand the inherent characteristics of a given text** and to adapt it accordingly.

Reference

- Biomedical Text Mining: State-of-the-Art, Open Problems and Future Challenges, Springer, January 2014
- Kocbek, S., Cavedon, L., Martinez, D., Bain, C., Mac Manus, C., Haffari, G., ... & Verspoor, K. (2016). Text mining electronic hospital records to automatically classify admissions against disease: measuring the impact of linking data sources. Journal of biomedical informatics, 64, 158-167.
- Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data processing and text mining technologies on electronic medical records: a review. Journal of healthcare engineering, 2018.
- Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support?. Journal of biomedical informatics, 42(5), 760-772.
- Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2017). Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE journal of biomedical and health informatics, 22(5), 1589-1604.
- Zhang, Y., Tiryaki, F., Jiang, M., & Xu, H. (2019). Parsing clinical text using the state-of-the-art deep learning based parsers: a systematic comparison. BMC medical informatics and decision making, 19(3), 77.
- Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., ... & Aerts, H. J. (2019). Deep learning predicts lung cancer treatment response from serial medical imaging. Clinical Cancer Research, 25(11), 3266-3275.
- http://openminted.eu/1129-2/
- Min Song, Keun Yong Kang, and Tatsawan Timakum, Examining Influential Factors for Acknowledgements Classification Using Supervised Learning *[Under-review]*


Thank you
Q & A