Effect of Preprocessing for Distributed Representations: Case Study of Japanese Radiology Reports

Taro Tada and Kazuhide Yamamoto
Nagaoka University of Technology
Nagaoka, Niigata, Japan
{*tada, yamamoto*}*@jnlp.org*

***Abstract*—A radiology report is a medical document based on an examination image in a hospital. However, the preparation of this report is a burden on busy physicians. To support them, a retrieval system of past documents to prepare radiology reports is required. In recent years, distributed representation has been used in various NLP tasks and its usefulness has been demonstrated. However, there is not much research about Japanese medical documents that use distributed representations. In this study, we investigate preprocessing on a retrieval system with a distributed representation of the radiology report, as a first step. As a result, we confirmed that in word segmentation using Morphological analyzer and dictionaries, medical terms in radiology reports are not handled as long nouns, but are more effective as shorter nouns like subwords. We also confirmed that text segmentation by SentencePiece to obtain sentence distributed representation reflects more sentence characteristics. Furthermore, by removing some phrases from the radiology report based on frequency, we were able to reflect the characteristics of the document and avoid unnecessary high similarity between documents. It was confirmed that preprocessing was effective in this task.**

***Keywords*-Japanese; Medical document; Radiology report; Distributed representation; Pre-processing;**

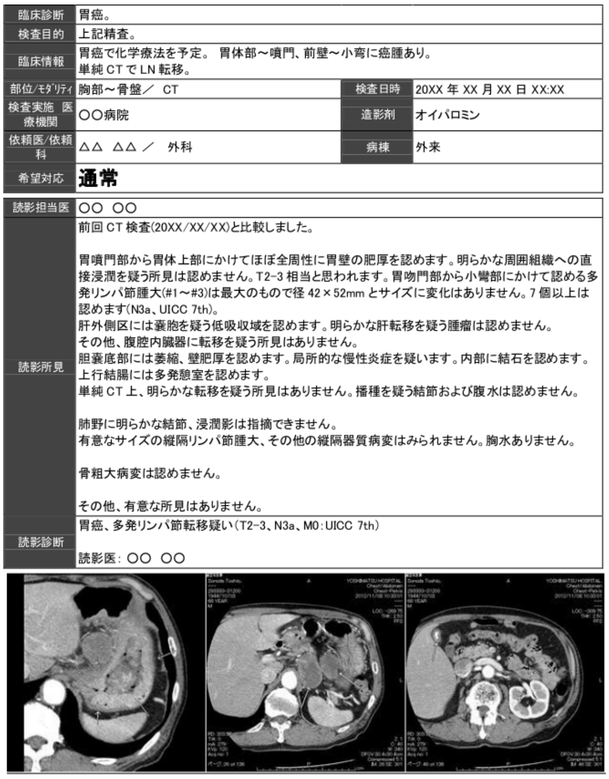
I. INTRODUCTION

Distributed representation has shown its usefulness in various tasks in the natural language processing field. Moreover, there are many studies on distributed representations in the biomedical domain [1] [2] [3]. However, in Japanese medical documents, distributed representations are used for some individual tasks, and there are not a very large number of studies as a whole. In Japan, attempts to digitize medical documents and their secondary use are underway; access to medical documents is generally difficult. Moreover, there are almost no medical domain data sets or corpus in the Japanese language that can be easily accessed by the general public, such as PubMed[1].

In this study, we used anonymous radiology reports. A radiology report is a report created after image examinations at hospitals. However, preparing this report is a burden for doctors, and they require similar past reports as a reference that include images and diagnosis.

Although there are similar document retrieval research works in the medical domain in Japanese, they have not been actively reported. Furthermore, there are only a few reports using distributed representation. Therefore, it is not known how to handle medical terms in Japanese, where it is necessary to perform word segmentation for tokenization. In addition, it is also unclear what kind of preprocessing is suitable because there are few studies on the basis of Japanese radiology reports.

In this study, we investigate the preprocessing for a similar document search in radiology reports using a simple method. We observed a tendency to obtain a better distributed representation in medical documents, especially in interpretation reports, by generating each document vector with the distributed representation method and evaluating the similarity between the documents.



臨床診断	胃癌。		
検査目的	上記精査。		
臨床情報	胃癌で化学療法を予定。 胃体部～噴門、前壁～小弯に癌腫あり。 単純 CT で LN 転移。		
部位/ﾓﾀﾞﾘﾃｨ	胸部～骨盤／ CT	検査日時	20XX 年 XX 月 XX 日 XX:XX
検査実施 医療機関	○○病院	造影剤	オイパロミン
依頼医/依頼科	△△ △△ ／ 外科	病棟	外来
希望対応	**通常**		
読影担当医	○○ ○○		
読影所見	前回 CT 検査(20XX/XX/XX)と比較しました。 胃噴門部から胃体上部にかけてほぼ全周性に胃壁の肥厚を認めます。明らかな周囲組織への直接浸潤を疑う所見は認めません。T2-3 相当と思われます。胃吻門部から小彎部にかけて認める多発リンパ節腫大(#1～#3)は最大のもので径 42×52mm とサイズに変化はありません。7 個以上は認めます(N3a、UICC 7th)。 肝外側区には嚢胞を疑う低吸収域を認めます。明らかな肝転移を疑う腫瘤は認めません。 その他、腹腔内臓器に転移を疑う所見はありません。 胆嚢底部には萎縮、壁肥厚を認めます。局所的な慢性炎症を疑います。内部に結石を認めます。 上行結腸には多発憩室を認めます。 単純 CT 上、明らかな転移を疑う所見はありません。播種を疑う結節および腹水は認めません。 肺野に明らかな結節、浸潤影は指摘できません。 有意なサイズの縦隔リンパ節腫大、その他の縦隔器質病変はみられません。胸水ありません。 骨粗大病変は認めません。 その他、有意な所見はありません。		
読影診断	胃癌、多発リンパ節転移疑い（T2-3、N3a、M0：UICC 7th） 読影医：○○ ○○		

Figure 1. An example of Japanese radiology report [4]

II. JAPANESE RADIOLOGY REPORTS

The medical documents dealt with in this study are anonymized radiology reports and examination order documents shared from Y's Reading inc. An examination order document is a document in which a doctor who examined a patient describes the findings and the purpose

[1] https://www.ncbi.nlm.nih.gov/pubmed/

of an examination of the patient's symptoms. An examination department, such as the radiology department, in a hospital, performs an examination based on the examination order document, and a doctor from the examination department describes the findings and diagnoses based on the photographed image. [4, Figure 1] shows an example of a radiology report.

There are two types of radiology report: primary and secondary radiology reports. The primary radiology report is written in reference to the examination image. The secondary report is created with reference to the primary report and the examination image. The secondary report, therefore, provides double-checking of radiology images together with the primary report. After the secondary report is prepared, it is submitted to the patient's doctor, and the doctor makes the final diagnosis. These three types of documents are written based on free description.

In departments such as radiology department that perform image examinations, the number of examinations per day is large, and in the case of examinations through CT (computed tomography) scan, the number of examination images per process may be very large, even several hundred. The preparation of such radiology reports is a burden on the doctors [5]. Therefore, in order to reduce the burden on the preparation of Japanese radiology reports, we investigate the preprocessing of a similar document retrieval system using distributed representations.

III. Related works

Research on natural language processing in medical domain documents, such as electronic medical records, there are many available resources, and many studies are presented at various conferences in English [6]. Moreover, there are many studies on radiology reports in English [7] [8]. However, studies using Japanese medical domain documents are not as frequent as those in English documents.

Research on retrieval systems for Japanese medical domain documents has been conducted before, however, there are not many. Ono et al. [9], Okamoto et al. [10], Doi et al. [11], Aramaki et al. [12] are representative. This is attributed to the fact that medical document data itself is difficult to obtain, and it is not so long after the secondary usage activities began [13]. In addition, the need for data anonymization, understanding the purpose and features of documents, and the appearance of domain-specific terminology also makes it difficult to handle documents without the cooperation of medical workers. For these reasons, the treatment of medical documents is complicated, which is considered to be a barrier in its research.

As prior research of similar document search in the medical domain, Okamoto et al. and Doi et al. searched for similar documents using TF-IDF. Okamoto et al. classified information of Observation, Diagnosis, and Treatment that are added to the sentences in medical records and are applied for similar document search. Doi et al. reported searching similar documents using a document vector space method based on TF-IDF. Ono et al. created a dendrogram of a maximum spanning tree from word weighting by TF-IDF and the ICD (International Classification of Diseases / International Statistical Classification of Diseases and Related Health Problems) code applied for classification of hospital discharge summaries. Aramaki et al. has been operating and reporting a disease case search system that performs a full-text search based on GETA(Generic Engine for Transposable Association), which is an associative search engine, for seven years.

Siddharth et al. [14] proposed a text preprocessing, that reduces the complexity of sentences in biomedical abstracts to improve the performance of syntactic parsers in English.

IV. Data

The data used are text data of 181,875 examinations. A examination order document, a primary radiology report, and a secondary radiology report are contained for each examination. Table I shows the statistical information of the radiology report and examination order document.

Table I
STATISTICAL INFORMATION OF RADIOLOGY REPORTS AND EXAMINATION ORDER

Type of document	Average document length (char.)
examination order	99.1
primary radiology report	299.3
secondary radiology report	332.3

In the radiology reports, the diagnosis is described for all the parts appearing in the photographed examination image. Suspected positive diseases and injuries will be described even if it is not the original purpose of the imaging examination. Radiology reports with single diseases are few, and those with multiple diseases appear in combination under findings and diagnoses in the report. Even if observing different diseases are the main purpose of the examination, there are cases wherein the same diseases are observed in reports. Furthermore, it is difficult to determine the degree of similarity between each disease. Therefore, it is difficult to judge the superiority or inferiority of the document similarity unless it is handled by a person with advanced medical knowledge, such as a doctor. Moreover, as per our knowledge, there is no data set on the Japanese radiology report that can be used to evaluate document similarity.

In this study, as a first step, we created a data set to simplify the problem. We investigated the effects of preprocessing by checking whether radiology reports included the same disease can be obtained as similar documents. The evaluation data set was created with reference to the study by Okamoto et al. [10].

First, we selected eight diseases (Alzheimer's disease, lung cancer, myocardial infarction, fatty liver, disc herniation, medial collateral ligament injury, Elbow fracture, Achilles tendon injury) as classes. Next, we selected radiology reports with selected the diseases in the above-mentioned. In selecting documents, different classes do not contain the same disease.

A total of 80 documents were selected, 10 documents in each disease class, as data for evaluation, although it is a small sample set. The documents for the 181, 795 examinations were used as training data, excluding the examination documents used for the evaluation data.

V. Preprocessing

We investigated the effect of preprocessing on the similarity between the documents. This process involved, first, word segmentation with different granularity for radiology reports; second, cleaning by removing frequently appearing phrases from the evaluation data.

Because the Japanese language is not described with a space between words, word segmentation is required as preprocessing for input to various models. These segmented words become system inputs, and the granularity of the segmentation may affect subsequent processes. In addition, as mentioned above, there are few studies on Japanese radiology reports; therefore, we investigated the effects of phrase frequency-based cleaning. The neologdn[2] was used to normalize characters before performing the above procedure.

A. Word segmentation

To investigate the effect of word segmentation with different granularities, word segmentation was performed using Japanese morpheme analyzers MeCab [15] and SentencePiece [16]. MeCab[3] can select a dictionary for use in word segmentation.

In this article, we used five dictionaries, including two medical terminology dictionaries. We used UniDic[4], IPADic[5], and ipadic-neologd[6] as general domain dictionaries, and ComeJisyo[7] and MANBYO Dictionary[8] as medical terminology dictionaries. These medical terminology dictionaries are combined with IPADic to supplement non-medical terminologies. Table II lists the examples of diseases that have been word segmented using MeCab with each dictionary. The medical terminology dictionaries can medical domain terminology to be treated as words with a relatively longer granularity than general domain dictionaries. The general domain dictionaries divide the medical terms into shorter granularities.

B. Word segmentation using SentencePiece

SentencePiece[9] can perform frequency-based segmentation, which is different from the dictionary-based word segmentation in terms of granularity. Therefore, we investigated the effect of segmentation using SentencePiece. Because SentencePiece can split text by specifying the number of unique tokens, we performed experiments on multiple unique token sizes. Other SentencePiece parameters were used as default.

Table II
Example of word segmentation difference by MeCab with dictionaries

dictionary	segmentation results
example1：右肺上葉切除術後	(After upper right lung lobectomy)
UniDic	右肺　上葉　切除　術後
IPADic,IPADic-NEologd	右　肺　上　葉　切除　術　後
MANBYO Dictionary, ComeJisyo	右　肺上葉切除　術　後
example2：陳旧性心筋梗塞	(old myocardial infarction)
UniDic	陳旧　性　心筋　梗塞
IPADic,IPADic-NEologd	陳　旧　性　心筋梗塞
MANBYO Dictionary, ComeJisyo	陳旧性心筋梗塞

C. Removing high-frequency phrases

As described in Section II, as a feature of the radiology report, the sentences of the findings are freely described in the radiology report for all parts that appear in examination images, even if they are not required for the original examination purposes.

In the radiology report, most of the findings and diagnoses of the areas not intended for examination are negative. These are written as sentences with body parts appear continuously (examples: no significant findings are found in liver, bile, pancreas, spleen) or semi-fixed sentences (example: no other significant findings). Depending on the body parts of the examination target and examination method, the sentences will appear in reports. In addition, because the radiology report is a report of an imaging examination, fixed sentences (such as greetings to the report receiver on the beginning or end of the finding) that are not related to the findings and diagnoses also appear.

As a result, it is assumed that the similarity unnecessarily increases when calculating the similarity between documents. Therefore, we checked for high-frequency phrases from the training data and manually selected phrases that were noisy and removed them from the evaluation data. Table III shows an example of the phrases removed from the evaluation data set.

Table III
Example of removed phrases

胆、(bile,) / 肝、(liver,) /その他、(other,) /
明らかな異常を認めません。(There is no obvious abnormality.) /
胸水みられません。(no pleural effusion)/
有意な所見はありません。(no significant findings.)/
ご報告申し上げます。(I'd like to inform you.) /
前回同様です。(same as last time.) /

VI. Experiments

Apply the preprocessing described in Section V to the training data and the collected evaluation data. A document vector is generated from each document of the evaluation data, and the similarity between each document is calculated to check the effect of preprocessing.

For each document vector of the evaluation data, the degree of similarity between the document vectors of the

[2] https://github.com/ikegami-yukino/neologdn
[3] https://taku910.github.io/mecab/
[4] https://unidic.ninjal.ac.jp/
[5] https://taku910.github.io/mecab/
[6] https://github.com/neologd/mecab-ipadic-neologd
[7] https://ja.osdn.net/projects/comedic/
[8] http://sociocom.jp/~data/2018-manbyo/index.html
[9] https://github.com/google/sentencepiece

remaining 79 documents in the evaluation data is calculated. Then, it is determined how many of the remaining nine documents in the same disease class are included in the top nine similar documents. This is performed for all the documents of the evaluation data. As a distributed representation method, a document vector is created using fastText [17], [18] with skip-gram, and the degree of similarity between the documents is calculated using cosine similarity.

fastText is the original implementation[10] by Facebook Inc. for investigating the effect of word segmentation by MeCab with different dictionaries; Doc2Vec [19] with DBoW was also employed. Doc2Vec is implemented with gensim[11]. The hyperparameter of Doc2Vec was determined with reference to Lau et al. [20] [21]. Table IV shows the hyperparameters during learning.

Table IV
HYPER PARAMETERS

	Hyper parameters (other parameters are default)
Doc2Vec	Method : DBoW, Dim : 300 Window Size : 30, Min Count : 5 Sub-sampling : 10^{-5} Negative sampling : 10, Epoch : 1000
fastText	Method : skipgram, Dim : 300 Window Size : 30, Min Count : 1 Negative sampling : 10, Epoch : 30 Loss : hs

VII. Results and Discussion

Table V shows the statistical information of the radiology report and examination order document after word segmentation, and Table VI shows the experimental results using MeCab and five different dictionaries. Unexpectedly, the accuracy of the results was improved by segmenting long medical terms into shorter words. This can be assumed because the short granularity of words reduced the number of vocabularies and increased the frequency of words by treating the medical terms as a group of words with finer granularity, like English subwords. However, single-character segmentation results in lower accuracy. In this case, even if each Kanji character has a meaning, it is likely that the meaning can be retained more by learning with short words. A similar trend was observed in fastText and Doc2Vec. These results may depend on the amount of data, but the data used in this experiment are relatively large compared to the actual Japanese medical texts available at present. Therefore, similar trends are likely to be obtained in many cases. On this result, the accuracy of fastText being lower than the that of Doc2Vec, because there was not any hyper-parameter tuning done of fastText. therefore accuracies of between methods are not comparable.

The effect of text segmentation by SentencePiece is shown in Table VII. The number of unique tokens of SentencePiece was the highest result in 4000, exceeding the accuracy of experimental results with MeCab and dictionaries. Some of the tokens in the documents that were divided using SentencePiece are phrases. Therefore, if medical terms are required to, for example, confirm the similarities between diseases during postprocessing, it is better to divide using MeCab with dictionaries. However, for obtaining a vector of a document, a better distributed representation can be obtained through dictionary-based word segmentation.

Table VIII shows the effect of cleaning, in which high-frequency phrases obtained from the learning data are removed from the evaluation data. As a result, it was confirmed that cleaning by phrase removing based on frequency and little domain knowledge was effective. In the case of medical documents with some fixed elements, high-frequency phrases unnecessarily increase the similarity between documents. We confirmed that removing of these phrases led to the acquisition of distributed representations that better reflect the characteristics of the documents.

In the case where documents of the same disease class were not obtained as similar documents in the experiment, some diseases were not included in the disease classes in the evaluation data and they unintentionally appeared over multiple classes. However, a high degree of similarity between the same disease documents is desirable, and in this respect, it can be said that the document vector reflects the contents of the document more accurately. In addition, based on the fact that they are influenced by words such as ”soft tissue” and ”fracture” that commonly appear when describing the findings of injuries and diseases, these documents can be said to be similar in a broad sense but require a different approach.

VIII. Conclusion

In this study, we investigated the preprocessing method for using distributed representations method for Japanese radiology reports, the effect of word segmentation granularity and cleaning based on frequently appearing phrases.

Preprocessing can have a great effect on obtaining distributed representations from radiology reports. Unexpectedly, it was more effective to divide the medical terms appearing in the radiology report into short words rather than being treated as long words, using a medical terminology dictionary. Moreover, the accuracy in the experimental result with word segmentation by SentencePiece is higher than morphological analyzer. It can be lead to consider that frequencies of words in the training data are very important for the learning of distributed representations better.

For cleaning the evaluation data, we used high-frequency phrases in the training data and selected high-frequency phrases based on a little domain knowledge and a little manual labor. We confirmed that this method could remove factors that unnecessarily increase the degree of similarity between documents, and reflect more document characteristics to the distributed representations.

The preprocessing in this study is common in many tasks using distributed representations, therefore it can be

[10] https://fasttext.cc/

[11] https://radimrehurek.com/gensim/index.html

Table V
STATISTICAL INFORMATION OF RADIOLOGY REPORTS AND EXAMINATION ORDER

Segmentation method	Average document length (word)			Average token length (character)		
	primary reports	secondary reports	examination orders	primary reports	secondary reports	examination orders
character segmentation	299.3	332.3	99.1	1	1	1
UniDic	212.2	237.0	68.2	1.41	1.40	1.45
IPADic	210.5	234.9	63.7	1.42	1.41	1.56
IPADic-neologd	202.9	226.7	60.2	1.48	1.47	1.65
MANBYO Dictionary	195.9	219.3	60.2	1.53	1.52	1.65
ComeJisyo	188.9	211.9	57.5	1.58	1.57	1.72
SentencePiece(4000)	132.8	147.6	59.4	2.25	2.25	1.67

Table VI
ACCURACIES OF FASTTEXT AND DOC2VEC IN CHANGING SEGMENTATION METHODS

segmentation method	fastText	Doc2Vec
character segmentation	0.685	0.799
UniDic	**0.708**	0.894
IPADic	0.704	**0.908**
ipadic-neologd	0.663	0.896
MANBYO Dictionary	0.644	0.828
ComeJisyo	0.613	0.869

Table VII
ACCURACIES OF FASTTEXT WITH SENTENCEPIECE SEGMENTATION

unique tokens	fastText	unique tokens	fastText
3000	0.836	16000	0.786
4000	**0.847**	18000	0.797
5000	0.824	20000	0.797
6000	0.832	25000	0.792
8000	0.819	30000	0.799
10000	0.772	40000	0.807
12000	0.769	100000	0.814
14000	0.781		

applied to various tasks.

IX. FUTURE WORK

Medical documents we used in this work were examination order documents, primary radiology reports, and secondary radiology reports. We confirmed a problem in a part of the word-level similarity (”handball” as upper similar words of ”metastasis”, etc.) that due to the influence of examination order documents. Probably, better distributed representations can be obtained by filtering the documents used for learning. Besides, it might be possible to learn relationships between medical terms using external documents.

In this study, we simplified the problem as the first step toward the similar document retrieval task of radiology reports. However, owing to the time and cost, the scale of the evaluation data was small. Because larger-scale evaluation data are required for considering overfitting risk. therefore we would like to create a new data set. In addition, we intend to collaborate with medical professionals with advanced medical knowledge in order to set up tasks closer to medical sites.

ACKNOWLEDGMENT

The authors are grateful to Y’s Reading inc. for giving incisive comments and sharing anonymized text data of examination documents.

Table VIII
EFFECT OF PHRASE REMOVING (ACCURACY)

Method	fastText	Doc2Vec	Avg. length (words)
IPADic	0.704	0.908	151.4
IPADic with removing	**0.781**	**0.922**	130.7

This work was supported in part by JSPS KAKENHI Grants-in-Aid for Challenging Research (Exploratory) Grant ID 17K18481.

REFERENCES

[1] Dieter Galea,Ivan Laponogov,Kirill Veselkov,Sub-word information in pre-trained biomedical word representations: evaluation and hyper-parameter optimization,Proceedings of the BioNLP 2018 workshop, pp.56-66 2018

[2] Chen, Qingyu and Peng, Yifan and lu, Zhiyong BioSentVec: creating sentence embeddings for biomedical texts, he Seventh IEEE International Conference on Healthcare Informatics (ICHI 2019),2019

[3] Alsentzer Emily, Murphy John, Boag William, Weng Wei-Hung, Jindi Di, Naumann Tristan, McDermott Matthew, Publicly Available Clinical BERT Embeddings, Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp.72-78 2019

[4] Y’s Reading, Our reports, Accessed on: Jul. 1, 2019. [Online]. Available https://ys-reporting.com/reports/

[5] Kyoko Makino, Rumi Hayakawa, Koichi Terai, Hiroshi Fukatsu, Development and Evaluation of a Diagnostic Documentation Support System using Knowledge Processing, Transactions of the Japanese Society for Artificial Intelligence, Vol.23 No.6 pp.463-472 2008

[6] Pranjul Yadav, Michael Steinbach, Vipin Kumar, Gyorgy Simon, Mining Electronic Health Records (EHR): A Survey, ACM Computing Surveys, Vol. 1, No. 1, Article 1, Publication date: April 2016.

[7] Banerjee I,Madhavan S,Goldman RE,Rubin DL,Intelligent Word Embeddings of Free-Text Radiology Reports, AMIA Annu Symp Proc. pp.411420 2017,Published 2018 Apr 16.

[8] Yifan Peng, Ke Yan, Veit Sandfort, Ronald M. Summers, Zhiyong Lu, ”A self-attention based deep learning method for lesion attribute detection from CT reports,” IEEE International Conference on Healthcare Informatics (ICHI), 2019

[9] Hiroki ONO, Katsuhiko TAKABAYASHI, Takahiro SUZUKI, Hideto YOKOI, Atsushi IMIYA, Youichi SATOMURA, Classification of Discharge Summaries by Text Mining, Japan Journal of Medical Informatics, Vol.24 No.1 pp.35-44 2004

[10] Kazuya OKAMOTO and Tadamasa TAKEMURA and Tomohiro KURODA, Keisuke NAGASE and Hiroyuki YOSHIHARA, Context-based Retrieval System for Similar Medical Practice Documents,Transactions of Japanese Society for Medical and Biological Engineering, Transactions of Japanese Society for Medical and Biological Engineering,Vol.49 No.6 pp.199-206 2006

[11] Shunsuke DOI, Takashi KIMURA, Masaki SEKINE, Takahiro SUZUKI, Katsuhiko TAKABAYASHI, Toshiyo TAMURA, Management and Evaluation of Similar Case Searching System in Homepage of Medical Society, Transactions of Japanese Society for Medical and Biological Engineering,Vol.49 No. 6 pp. 870-876 2011

[12] Aramaki E, Iwao T, Wakamiya S, Ito K, Yano K, Ohe K, A Fundamental Study on User Utilization Based on a Trial Operation of the Medical Case Retrieval System, Japan Journal of Medical Informatics, Vol.38 No.4 pp. 245-256 2018

[13] Toshihiro TAKEDA and Shirou MANABE and Yasushi MATSUMURA, The Current Situation and Issues of the Secondary Use of Electronic Medical Record Data, Transactions of Japanese Society for Medical and Biological Engineering,Vol.55 No.4 pp. 151-158 2017

[14] Jonnalagadda Siddhartha, Tari Luis, Hakenberg Jörg, Baral Chitta, Gonzalez Graciela, Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text, Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp.177-180 2009

[15] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto, Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp.230-237 2004

[16] Taku Kudo and John Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (System Demonstrations), pp.66-71 2018

[17] Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov, Enriching Word Vectors with Subword Information, Transactions of the Association for Computational Linguistics, vol.5 pp. 135-146 2017

[18] Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov, Bag of Tricks for Efficient Text Classification, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2, Short Papers, pp. 427-431 2017

[19] Quoc Le, Tomas Mikolov, Distributed Representations of Sentences and Documents, Proceedings of the 31 st International Conference on Machine Learning, JMLR: W&CP volume 32, pp. 1188-1196 ,2014

[20] Jey Han Lau and Timothy Baldwin, An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation, Proceedings of the 1st Workshop on Representation Learning for NLP, pp. 78-86 2016

[21] Billy Chiu,Gamal Crichton,Anna Korhonen,Sampo Pyysalo, How to Train Good Word Embeddings for Biomedical NLP, Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pp.166174 2016