International Journal of Advance in Computer Science & its application – IJCSIA 2018
Copyright © Institute of Research Engineers and Doctors , SEEK Digital Library
Volume 7 : Issue 2 [ISSN : 2250-3765] - Publication Date: 25 June , 2018

Interactive Visualization of Template Graph for Daily Clinical Notes*

N. Onimura[1)], T. Yamashita[2)], N. Nakashima[2)], H. Soejima[3)], and S. Hirokawa[4)]

***Abstract*— Clinical notes in medical institutions are essential not only for information sharing among medical staff but also\ for analysis to improve medical activities. Especially, the record of free description is important for grasping the situation of the patients. However, manually recording has the problem of input labor and fluctuation of expression by the person who records. In order to solve this problem and enable statistical analysis and mechanical analysis, templates are expected.**

This paper proposes an interactive support system for defining medical record template for each disease by extracting feature words corresponding to the disease from medical records using machine learning. We constructed an interactive system that displays the context of those characteristic words as a directed graph. A pull-down template can be generated from this graph. We developed a system to generate such graphs interactively for 123,736 free description clinical records of a hospital and examined its usefulness.

I. INTRODUCTION

In various fields, hand written records are being replaced by digitalized ones and are being used for further analysis to improvement the activities. Detailed description can be possible, they are kept in free description format. Particularly, in medical institutes, recording of free description has advantages that it records the of situation of individual patients and can be shared by other stuff. However, the diversity of free description sentences may cause ambiguity and difficulty, and can not be quantitatively analyzed. In order to realize the quantitative analysis of the document records, it is necessary to eliminate the diversity of expressions and to reduce the burden of the entry. To make templates for free description sentences is one of ideal solutions to this problem.

In order to create a template depending on each disease, the present paper proposes a method to generate directed graphs as a visualization of typical clinical notes for each disease. We realized the method as an interactive system by which doctors and nurses can construct templates as they want. Firstly, we constructed a search engine for 123,736 clinical notes on "pain" collected from April 2014 to August 2016 at Saiseikai Kaikan Hospital. The targeted clinical notes are free description documents described in a fixed form for each clinical path. The clinical path defines procedures for examination and treatment according to the disease and the condition of the patient. Every doctor or nurse does not decide arbitrarily how to inspect, medicate, diagnose, and treat patients. Procedures according to each disease are decided as organization. The description is not a completely free form but a semi-structured document called SOAP form which consists of Subjective, Objective, Assessment and Plan [1].

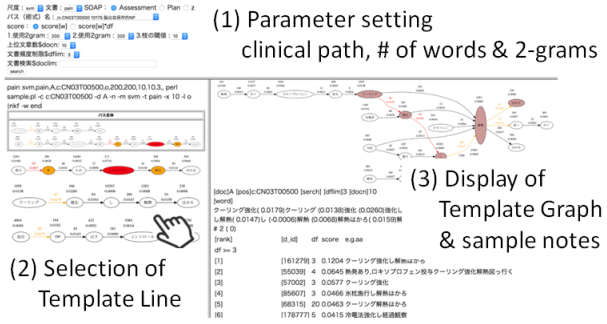


Figure 1 Snapshot of Interactive System

The final goal of this study is to generate templates for clinical path compliant medical records. What is needed is not a template for general practice activities, but a different template according to each disease. Automatic generation of a template corresponding to disease is a challenging theme. However, in order to use it in an actual hospital, judgment of a responsible doctor and medical stuff is crucial. The templates should be based on real medical records and must be consistent with their experience. Therefore, we construct a system for supporting the formulation of templates, instead of a fully automatic template generation system.

Firstly, the system extracts feature words according to each disease, i.e. according to each clinical path, and then the system displays the characteristic contexts in which those words are used. Doctors can check the typical sentences that are related to those contexts. Figure 1 is a screen shot of the interactive system. A doctor or anurse can specify a clinical path in the part (1) of Figure 1. Then, the characteristic words are extracted and the template graphs are shown in the part (2). If he/shes select a graph in (2), then, the corresponding sentences are shown in the lower part of (3) of the right frame and a merged graph of those sentences are shown in the upper part (3). In this system, multiple parameters can be specified, and graphs of various patterns can be generated. By interactively repeating the process of parameter selection, graph generation, and sentence example confirmation by trial and error, it is possible to create a satisfactory graph.

* Research supported by Grants-in-Aid for Scientific Research (KAKENHI (Multi-year Fund))15H02778.

1) N. Onimura is with the Graduate School of Information Science and Electrical Engineering, Kyushu University, Motooka 744, Fukuoka 819-0395, JAPAN.

2) T. Yamashita and N. Nakashima are with Medical Information Center, Kyushu University Hospital, 3-1-1 Maidashi Higashi-ku Fukuoka, JAPAN ({t-yama,nnaoki}@med.kyushu-u.ac.jp).

3) H. Soejima is with Saiseikai Kumamoto Hospital, 5-3-1 Chikami Minami-ku Kumamoto city Kumamoto 861-4193, JAPAN (hidehisa-soejimapsaiseikaikumamoto.jp).

4) S. Hirokawa is with Research Institute for Infomation Technology, Kyushu University, Motooka 744, Fukuoka 819-0395, JAPAN (corresponding author, hirokawa@cc.kyushu-u.ac.jp)

II. OVERVIEW OF THE INTERRACTIVE SYSTEM

The basic ideas from document vectorization to feature extraction and context visualization are as follows. For vectorization of 123,736 documents, we used not only words, but also word 2-gram that appears adjacent. All sentences are vectorized by single words, word 2-grams and these document frequencies. Figure 1 (1) shows the parameter choice for the clinical path of "cerebral hemorrhage".

In the next step, we apply the machine learning method SVM (support vector machine) with 9595 clinical notes on the patients of "cerebral hemorrhage" as positive data, and the other 114141 clinical notes as negative data. SVM generates a model to predict if a document is positive or negative according to the weight of each word. In our formulation, not only words but also word 2-grams are used as feature. Thus, we can tell which words and word 2-grams are important in distinguishing "cerebral hemorrhage" patients and others. The higher the score, the closer it is to the disease.

The lower part (2) of Figure 1, displays "template lines" that contains characteristic 2-grams for "cerebral hemorrhage" that have highest score.

A click on a "template graph" (Figure 1. (1)) retrieves the all the sentences and displays the top 10, in this case, as specified in Figure 1 (1), at the lower part of the right frame Figure 2 (3). Those 10 sentences are displayed as a graph by merging the same word as the single node. We call such a graph as a template graph.

By learning data other than positive examples using SVM, we could extract characteristic words not dependent on document frequencies of simple words. Moreover, by using 2-gram, we made it possible to visualize the characteristic context which could not be done by only word vectorization.

III. RELATED WORK

Templates are typical data of semi-structured documents having properties intermediate between structured numerical data and freely-written sentences. Templates are also known to be useful for data transfer between different databases in medical institutions ([3, 4]). Not only for reuse but also for resolving ambiguities and trouble in expression when entering data, the effectiveness of common template is expected. Templates are an effective way to store and share records. Extracting common templates from medical records of free description, which have been accumulated in medical institutions so far, is an important issue that cannot be ignored in order to advance electronization of medical records.

Preceding research on template extraction have been mainly focused on Web pages [5, 6, 7, 8]. Web pages generated from databases and CMS (contents management system) contain routine patterns. Indeed, many pages contain such patterns [9]. In early studies, templates are extracted by analyzing the patterns of Web pages described in HTML. Template extraction targeting general documents other than Web pages is more difficult since there is no structural clue like HTML tags in Web pages. In many researches, templates are extracted by using natural language processing technology, limiting objects to be extracted using templates, related technical terms and co-occurring common words as cues. Chambers and Jurafsky [10] focused on the feature words such as places and people to detect terrorism. Chang et al. [11] describes extraction and use of templates representing emotion. Proskurniay et al. [12] performs template extraction for e-mails. There are research on template extraction from scientific articles, especially from medical literature [13, 14]. [15, 16, 17] are extracting important sentences in medical literature. Template extraction is considered as effective clue for discovering technical terms and for constructing ontology [18, 19, 20, 21, 22]. On the other way round, by using them as prior knowledge, it is expected to increase the efficiency of template extraction. There are studies to capture templates by obtaining the IMRAD structure of the paper (introduction, method, result, discussion) with the characteristic words used in each sentence [2, 23]. They are not limited to medical literature.

In many studies, templates are extracted by restricting terminology for specifying semantic of a target template and specifying syntax. This paper can also be regarded as research on acquisition of prior knowledge for template discovery. A similar approach is [7], where they firstly make clustering the documents and obtain feature words of each cluster and extract templates by using them as clues. On the other hand, the present paper applies classification and feature words extraction. Specifically, by restricting the purpose of clinical path, we label a medical document as a positive and negative, and apply machine learning to obtain characteristic words of positive examples, and we use them as template clues. Another feature of the present paper is the way to validate the obtained template. Many of the previous studies have been evaluated using experimental data and human-prepared correct answer data. The data handled in this study is actual data accumulated in a hospital on site rather than data for evaluation experiments. Therefore, for evaluation by on-site medical personnel in the field, we constructed a system that links templates to visualization as a graph and case examples of typical sentences. Some of the templates are already used to improve medical records.

IV. FEATURE EXTRACTION AND GRAPH GENERATION

A. Extraction of Characteristic 2-grams by SVM

SVM (Support Vector Machine) is a machine learning technique for constructing a model that classifies the data into the positive data and negative data according to the training data. SVM selects the optimal hyperplane that maximizes the margin. In the present paper, we used the “word 2-grams” instead of single words. A 2-gram is a pair (*wi, wj)* of two words that occur adjacently in a sentence. We used a morphological analysis tool Mecab to obtain 2-grams. By applying SVM, we evaluated how each 2-gram is crucial to distinguish the target sentence. In the present paper, the target sentences are determined by an operation name or by a

International Journal of Advance in Computer Science & its application – IJCSIA 2018
Copyright © Institute of Research Engineers and Doctors , SEEK Digital Library
Volume 7 : Issue 2 [ISSN : 2250-3765] - Publication Date: 25 June , 2018

1. Construct search engine with index of words and word 2-grams and retrieve the set $D(q)$ of sentences given an operation name or a pathway.
2. Apply SVM with $D(q)$ as positive data and construct the model and classify the imaginary sentence that consists of single 2-gram (w_i, w_j) using the model to get the score of the 2-gram.
3. Select the top N 2-grams based on this score.

Figure 2. Selection 2-grams

pathway. 2-grams is more powerful to extract characteristic sentences than co-occurrent words. We used the method of [Adachi2016] for evaluating the score of 2-grams. Given an operation name or a pathway, we extract characteristic 2-grams by the process shown in Fig. 2.

Given a set of documents for a query q, integers N and M, we generate a set of directed graphs $G(q, N, M)$. Actually, the set $G(q, N, M)$ consists of straight line graphs each of which contains a top N 2-grams. The score of 2-grams and the score words are determined by the SVM-model with respect to the query. In the algorithm, P denotes the set of graphs, E denotes the set of 2-grams, $n(P)$ denotes the total number of edges of P. Thus, the graph is expaned in tail direction as shown in Fig.3.

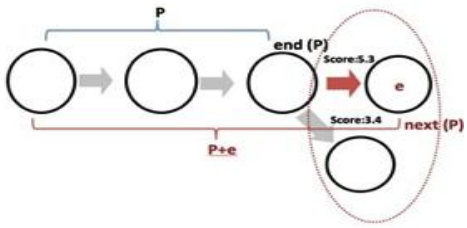


Figure 3. Tail Expansion

B. *Support Sentences of Template Graph*

Directed graphs are useful to grasp the patterns that appear among the sentences obtained as the search result. However, they are not convincing enough unless we see the real sentences that match the pattern. In this section, we describe a method that lists such sentences given a template graph. The basic idea is to consider a graph as a "sub-model" for the classification by SVM with respect to the query q. Let $G(q)$ be a template path obtained by the algorithm. We know the SVM-scores of words and 2-grams with respect to the SVM-model. Given a sentence s_i, we calculate the score $scr(s_i, G(q))$ as follows:

$$scr(s_i, G(q)) = \sum_{w_j \in G} tf(w_j, G(q)) * scr(w_j) + \sum_{u:v \in G} tf(u:v, G(q)) * scr(u:v)$$

, where w_j is a word that appears in the graph $G(q)$, $u:v$ is a 2-gram that corresponds to an edge of the graph $G(q)$, $scr(x)$ is the SVM-score of the word and the 2-gram with respect to

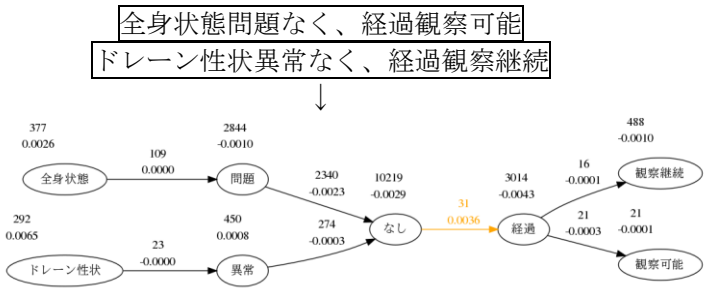


Figure 4. Example for Sentences Graph

the SVM-model and $tf(x, G(q))$ denotes the term frequency of the word and the 2-gram. The score $scr(s_i)$ of a sentence s_i is the sum of weighted score of those words and 2-grams. In order to deepen the visual interpretation of the list of documents, we visualized all 2-grams in sentence s_i as a directed graph using nodes and edges, in the same way as the above directed graph.

C. *Interactive Visualization*

We created an interactive visualization system which yields the directed graph or document list interactively while freely changing parameters. The following is the list of parameter that can be changed interactively.

Method : "SVM, GETA, …"
Analysis method of document
Target : "pain, …"
Document name (clinical case) to be analyzed
SOAP : "Assessment, Plan, …"
Assessment document or Plan document
Clinical pathway : "cerebral infarction, brain hemorrhage, …"
Clinical pathway or surgery for the case
Number of edges : "10, 15, 20, …"
Number of features to extract 2-gram
Number of sentences : "10, 15, 20, …"
Number of sentences to display
Limit of document frequency : "2, 3, 4, …"
Displays only sentences above the specified document frequency
Search word : "blood pressure, analgesics, follow-up observation, …"

V. DATA

We constructed a search engine for 123,736 pain variance records collected from April 2014 to August 2016 in Saiseikai Kumamoto hospital. We used word 2-grams that occur adjacently in a sentence of the given documents. All the sentences are vectorized by words and word 2-grams. Then we applied SVM to construct a model for classifying the sentences into positive ones and negative ones. We conducted this on two treatment policies, "brain hemorrhage" and "cerebral infarction". A total of four medical documents are described for treatment to stroke

International Journal of Advance in Computer Science & its application – IJCSIA 2018
Copyright © Institute of Research Engineers and Doctors , SEEK Digital Library
Volume 7 : Issue 2 [ISSN : 2250-3765] - Publication Date: 25 June , 2018

Table 1. TOP 4 CLINICAL PATHWAYS

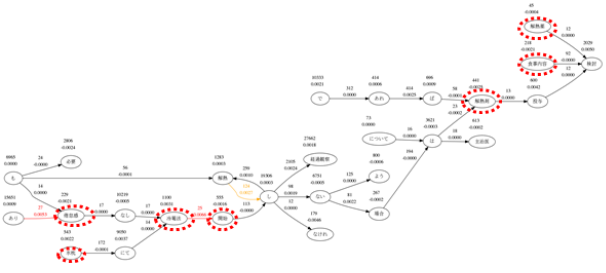
clinical pathway	count	assess	plan
blood purification	14512	11603	13407
brain hemorrhage	10175	6655	5769
cerebral infarction	9595	6097	5435
proximal femoral fractures	8507	1441	6882

VI. EVALUATION OF GRAPHS

The words appearing in the graph are extracted from the document written by doctors and nurses. They are medical documents of treatment plan and surgery. The characteristic words of the clinical treatment are highlighted with red circle. The visualization indicates the important sentences and phrase in the clinical treatment.

Brain hemorrhage (Assessment) graph shows words of fatigue, cooling, antipyretic and food (Fig. 4.1). Brain hemorrhage (Plan) graph shows words of analgesics, cooling and fever (Fig. 4.2). Cerebral infarction (Assessment) graph shows words of appetite, digestive system symptom, nervous system symptom, oxygen saturation, blood pressure and aspiration (Fig. 4.3). Cerebral infarction (Plan) graph shows words of analgesic and the administration (Fig. 4.4). We can see feature words or sentences related to stroke case. Specifically, they are cooling and taking antipyretic for decline of fever, taking analgesics for relieving pain, checking patient condition (blood pressure, aspiration, nervous system symptom). Fever is described to be associated with poor prognosis in acute stroke by previous study [24, 25, 26].

Particularly in the graph of cerebral hemorrhage, feature words or feature sentence related to the patient condition appear in the Assessment graph. Related words to the treatment plan appear in the Plan graph. In the graph of brain hemorrhage, similar words or sentences were extracted between the Assessment graph and the Plan graph.

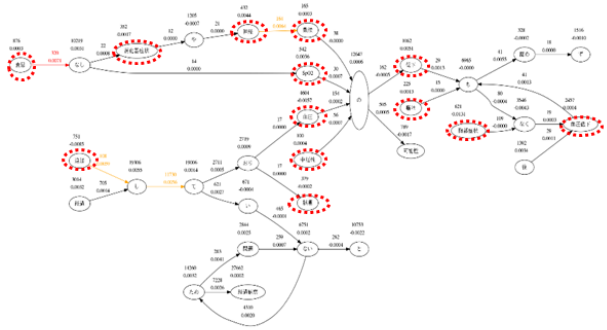


倦怠感 / [あり＊なし] fatigued / not fatigued
・氷枕/にて / 冷罨法 / 開始 Cooling with ice pack
・ 解熱 / しない / 場合 / は / 解熱剤 / [投与*検討]
If the fever does not subside, [take /consider] an antipyretic
・ 食事内容 / 検討 Consider food menu

Figure 5 Template Graph of "brain hemorhage"(Assessment)

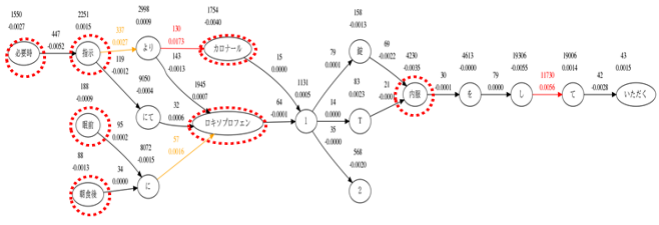
- 鎮痛薬 / 希望 / [あり＊なし] /氷枕 / 対応 Hope for analgesics [yes / no], and then take ice pack
- 鎮痛剤 / 内服 take analgesics
- 熱型 / 経過観察 For fever / observe progress

Figure 6 Templage Graph of "brain hemorhage"(Plan)



・食思 / [あり＊なし] Appetite [yes / no]
・消化器症状 Gastrointestinal symptoms
・神経兆候 / [悪化*変化] Neurosis / [worsening * changes]
・[SPO2＊血圧] / 低下 [SPO2 / blood pressure] / decline
・胸部症状 Chest symptom
・嘔吐 / 認める Vomiting
・[湿布＊食事] / 追加 [Poultice / dietary] add
・中枢性 / (発熱) Central / fever
・誤嚥 Aspiration

Figure 7 "cerebral infarction"(Assessment)



- 必要時 / 指示 / [カロナール＊ロキソプロフェン] / 内服
Direction when necessary / [Caronar * Loxoprofen (analgesic)] / intake per oral
- [眠前＊朝食後] / [カロナール＊ロキソプロフェン] / 内服
[Before sleeping / After breakfast] / [Caronar * Loxoprofen (analgesic)] / intake per oral

Figure 8 "cerebral infarction"(Assessment)

VII. CONCLUSION AND FURTHER WORK

In this paper, we constructed an interactive visualization system for feature extraction from document records. We can change the parameters to analyze exploratory. This system provides both the list of sentences and a sentence graph for deep visual interpretation. In medical documents, we extracted expressions of "common" as a directed graph which is particularly important when describing a medical document related to clinical plan and clinical treatment (surgery) for the clinical case. Specially, we extracted feature words (cooling and taking antipyretic) related to stroke treatment. Furthermore, in the graph of cerebral hemorrhage, words of the patient condition and words of the plan were expressed. We have verified these results can be considered as templates to hospital information system for improving medical practice. As further work, we would like to evaluate the effect of 2-grams. We also would like to introduce clustering mechanism for template graphs.

REFERENCE

1. L.L. Weed, Medical Records, Medical Education, and Pati ent care. Cleaveland, Western Reserve University, 1969.
2. Naoya Onimura, Takanori Yamashita, Naoki Nakajima, Hidehisa Soejima, Sachio Hirokawa, Generation of Sentence Template Graph from SOAP Format Medical Documents, Proc. CSCI2016, pp.159-162, 2016
3. Matsumura Yasushi, Murata Taizo, Horishima Hiroyuki, Ikebe Yoshie, Shimoji Ikuyo, Takeda Toshihiro, Hasegawa Hiroaki, Iwasaki Tetsuya, Watakabe Hiroyuki, Sharing and secondary utilization of data in the documents generated by multivendor systems, Proc. Joint Conference on Medical Informatics, pp.967-970, 2010 (in Japanese)
4. Yasushi Matsumura, Taizo Murata, Hiroyuki Horishima, Yoshie Ikebe, Ikuyo Shimoji, Toshihiro Takeda, Hiroaki Hasegawa, Tetsuya Iwasaki, Hiroyuki Watakabe, Sharing and secondary utilization of data in the documents generated by multivendor systems, Proc. Joint Conference on Medical Informatics, pp.967-970, 2010 (in Japanese)
5. Julian Alarte, David Insa, Josep Silva, Salvador Tamarit, TeMex: The Web Template Extractor, Proc. WWW2015, pp.155-158, 2015
6. Chulyun Kim, Kyuseok Shim, TEXT: Automatic Template Extraction from Heterogeneous Web Pages, IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 4, pp.612-626, 2011
7. Rashmi D Thakare, Manisha R Patil, Extraction of Template using Clustering from Heterogeneous Web Documents, International Journal of Computer Applications, Vol.119, No.11, pp.23-31, 2015
8. Neethu Mary Varghese, Tenny Thomas Soman, A Survey On Various Web Template Detection And Extraction Methods, International Journal of Scientific & Technology Research, Vol.4, Iss.3, pp.16-41, 2015
9. David Gibson, Kunal Punera, Andrew Tomkins, The Volume and Evolution of Web Page Templates, Proc. WWW2005, pp.830-839, 2005
10. Nathanael Chambers, Dan Jurafsky, Template-Based Information Extraction without the Templates, Proc. ACL2011 (Annual Meeting of the Association for ComputationalLinguistics: Human Language Technologies),Vol.1, pp.976-986, 2011
11. Yung-Chun Chang, Cen-Chieh Chen, Yu-Lun Hsieh, Chien Chin Chen, Wen-Lian Hsu, Linguistic Template Extraction for Recognizing Reader-Emotion and Emotional Resonance Writing Assistance, Proc. ACL2015, pp.775-780, 2015
12. Julia Proskurniay, Marc-Allen Cartrightz, Lluis Garcia-Pueyoz, Ivo Krkaz, James B. Wendtz, Tobias Kaufmannz, Balint Miklosz, Template Induction over Unstructured Email Corpora, Proc. WWW2017, pp.1521-1530, 2017
13. Fei Zhu, cheng Zhan, Yang Yang, Jonathan Chan, Asawin Meechai, Wanwipa vongsangnak, Bairong Shen, Biomedical text mining and its applications in cancer research Journal of Biomedical Informatics, Vol. 46, Iss.2, pp.200-211, 2013
14. Yuan Luo, Aliyah R Sohani, Ephraim P Hochberg, Peter Szolovits, Automatic lymphoma classification with sentence subgraph mining from pathology reports, Journal of the American Medical Informatics Association, 21(5), pp.824-832, 2014.
15. Anna Divoli, Teresa K. Attwood, BioIE: extracting informative sentences from the biomedical literature, Bioinformatics, Vol. 21, No.9, pp.2138-2139, 2005
16. Daigo Inoue, Hidetoshi Nagai, Teigo Nakamura, Hirosato Nomura, Harutosi Oogai, Analysing Description Patterns for Information Extraction from Abstracts of Medical Articles, IPSJ Technical Report, NL-141, No.9, pp.103-110, 2001 (in Japanese)
17. Satoshi Kamegai, Kenji Satou, Akihiko Konagaya, Automated Template Discovery for Information Extraction from Biomedical Literature, Proc. of the International Conference on Cybernetics and Information Technologies, Systems and Applications (CITSA 2004), Vol II, pp.39-44, 2004
18. Ken Yano, Kaoru Ito, Shoko Wakamiya, Eiji Aramaki, Development and Performance Evaluation of Deep Learning Method for Extraction of Named Entities from Medical Case Reports, Proc. Annual Conf. Japanese Sciety for Artificial Intelligence, 2017 (in Japanese)
19. Hiroshi Nakagawa, Automatic term recognition based on statistics of compound nouns, Terminology, Vol.6, No.2, pp.195-210, 2000
20. Taisuke Ogawa, Tomoyoshi Yamazaki, Ryo Sai, Mitsuru Ikeda, Muneou Suzuki, Kenji Araki, Koiti Hasida, Knowledge Sharing Support Based on Medical Service Ontology, Proc. Annual Conf. Japanese Society for Artifical Intelligence, 2009 (in Japanese)
21. Kazuhiko Ohe, Ken Imai, Development of Medical Ontology for Clinical Knowledge Processing, Journal of the Japanese Society for Artificial Intelligence, Vol.25, No. 4,pp.493-500, 2010 (in Japanese)
22. P. Turney, Learning to Extract Key Phrases from Text, Technical Report ERB-1057, NRC-41622, 1999
23. Yusuke Adachi, Naoya Onimura, Takanori Yamashita, Sachio Hirokawa, Standard Measure and SVM measure for Feature Selection and Their Performance Effect for Text Classification", Proc. iWAS2016, pp.262-266,2016
24. Koutarou Matsumoto, Nohara Yasunobu, Yoshifumi Wakata, Takanori Yamashita, Naoki Nakashima, Exploratory Data Analysis of Clinical Pathway for Brain Hemorrhage Using Machine Learning Technique, Proc. CJKMI2016, 2016.11.22

25. Hajat Cother, Hajat Shakoor, Sharma Pankaj, Effect of poststroke pyrexia on stroke outcome: a meta-analysis of studies in patients, Stroke ,Vol.31, No.2, pp.4104-414, 2000
26. Prasad Kameshwar, Krishan PR, Fever is associated with doubling of odds of short-term mortality in ischemic stroke: an updated meta-analysis, Acta Neurol Scand,Vol.122, No.6, pp.404-408, 2010