

Nursing-care Text Classification using Additional Term Information from Web

Manabu Nii^{*§}, Takafumi Yamaguchi^{*}, Yusuke Mori[†], Yutaka Takahashi^{*},
Atsuko Uchinuno[‡], and Reiko Sakashita[‡]

^{*}Graduate School of Engineering, University of Hyogo, Himeji, Hyogo, Japan
Email: nii@eng.u-hyogo.ac.jp

[†]School of Engineering, University of Hyogo, Himeji, Hyogo, Japan

[‡]College of Nursing Art & Science, University of Hyogo, Akashi, Hyogo, Japan

[§]WPI Immunology Frontier Research Center, Osaka University, Suita, Osaka, Japan

Abstract—In this paper, for improving performance of the nursing-care text classification, we introduce a mechanism of retrieving terms from Web. Every year, the nursing-care texts are collected by using Web application to improve nursing-care quality in Japan. The collected nursing-care texts are decomposed into morphemes (i.e., terms), and then terms are stored as a term list. Each text is represented as a feature vector by using the term list and classified using a SVM based classification system. The training data sets for constructing SVM based classification system are different from the evaluation data sets. That is, there are differences between the term lists of the nursing-care texts because the nursing-care texts are collected and evaluated every year. To cover this difference, we introduce a mechanism of retrieving terms from Web. A new term which appeared in the evaluation data sets is used as a query of a search engine. The terms in the term list are also used as queries. Terms are represented by the search results, and then are compared with each other. We use the most similar term in the term list as an alternative of the new term. From experimental results, we show effectiveness of our proposed method.

Keywords—nursing-care; text classification; SVM; fuzzy

I. INTRODUCTION

In order to improve nursing-care quality, the data which consists of numerical data and freestyle texts are stored into our database. The numerical data is used for statistical analysis about the structure of hospitals. The text data is used for evaluating each nurse. In evaluating the text data, the experts need to read carefully and evaluate each nursing-care data because those data include freestyle Japanese texts which are written about actual nursing-care by nurses. For evaluating such texts automatically, we have developed several nursing-care text classification systems, such as using fuzzy classification systems [1], neural networks [2], and support vector machines (SVM) [3], [4]. In order to improve the classification performance of the SVM based classification system, we have proposed a GA-based term selection [5]–[8].

In the above mentioned classification systems, the collected nursing-care texts are decomposed into morphemes (i.e., terms), and then terms are stored as a term list. Each text is represented as a feature vector by using the term list and classified using a classification system. The training data sets for constructing classification system (e.g., SVM based

classification system) are different from the evaluation data sets. That is, there are differences between the term lists of the nursing-care texts because the nursing-care texts are collected and evaluated every year. Therefore, the corresponding feature in the newer data set may not exist in the training data set because every year's term list has differences with each other. Such a feature which appears only in the new data set does not affect classification. The nursing-care texts which have such features may not be classified correctly.

In order to decrease such no-effect feature, we introduce a mechanism of retrieving terms from Web. A new term which appeared in the evaluation data sets is used as a query of a search engine. The terms in the term list are also used as queries. Terms are represented by the search results, and then are compared with each other. We use the most similar term in the term list as an alternative of the new term. Therefore, the no-effect feature can be expressed in the known feature approximately.

This paper is organized as follows. The overviews of nursing-care texts and preprocessing for classification using the morphological analysis software "MeCab" [9] are described in Section 2. In Section 3, the overview of nursing-care classification system and, in Section 4 our proposed method are explained. Then, simulation results are shown in Section 5.

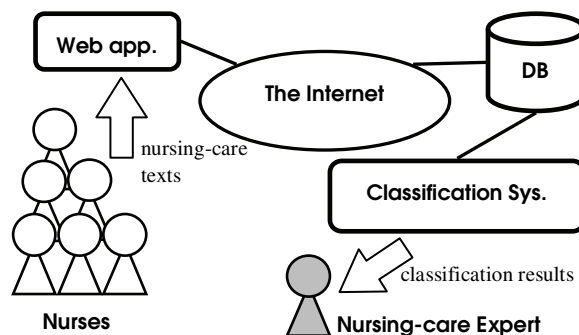


Fig. 1. Overview of our nursing-care data collection and classification

II. NURSING-CARE TEXT

The nursing-care text data set consists of many freestyle Japanese texts. They are collected by using an Internet application and stored into our database. The overview of nursing-care data collection and classification is shown in Fig.1.

The nursing-care texts contain answers for a wide variety of questions from nurses about their subjects, technicality, etc. Some sample texts are shown in Fig.2. Usually long texts are good texts which explained good nursing-care. On the other hand, usually short texts are poor and lacking the appropriate technical information. Some nursing-care experts read those texts in detail, then evaluate the nursing practices of the nurses providing the information, and classify them into classes 1 (bad) to 4 (very good).

In our nursing-care text classification system, already classified nursing-care texts by experts are used as a training data set for a SVM. In preprocessing, a text set is converted from original texts to a “term-document matrix” by decomposing it into morphemes. We use “McCab” as one of the morphological analysis software. All nouns and verbs (shown in shaded, see Fig.2) in decomposed morphemes are stored into a term list. We iterate the preprocessing until all texts are processed.

III. NURSING-CARE TEXT CLASSIFICATION SYSTEM

A. Making feature vectors

A term list which stored terms collected in a certain year is represented by TL_{year} . For example, a term list of 2008 is TL_{08} . Each nursing-care text is converted into a feature vector by assigning ‘1’ to that feature when a term in the nursing-care text existed in TL_{year} . Specifically, a nursing-care text is converted into a feature vector as the following manner.

Step 1 A nursing-care text is decomposed into morphemes.

Step 2 Nouns and verbs are selected from morphemes.

Step 3 If a selected term (i.e., a noun or verb) is appeared in the TL_{year} , then ‘1’ is assigned to that feature. Otherwise ‘0’ is assigned. That is, a feature vector is expressed as follows:

$$(x_1, \dots, x_i, \dots, x_{N_{TL_{year}}}), \quad (1)$$

$$x_i = \begin{cases} 1 & \text{if the term is appeared in } TL_{year}, \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where $N_{TL_{year}}$ is the number of terms in the term list TL_{year} .

Step 4 Till all texts are converted, repeat Step 1 to 3.

A SVM based classification system is constructed by using the feature vectors as a training data set.

B. Discussion

In this subsection, we show some experimental results and problems for the unseen nursing-care text classification. The nursing-care classification system has been developed to classify the nursing-care text automatically. That is, in actual use, the classification system is constructed by using the last year’s data set. The constructed classification system

TABLE I
CLASSIFICATION RESULTS ON THE 2008 DATA (10CV)

Data Index	# of correctly classified patterns
P123	329/569 (57.8%)
P131	361/519 (69.6%)
P132	175/400 (43.8%)
P212	322/481 (66.9%)
P213	281/509 (55.2%)
P221	341/512 (66.6%)
P222	382/483 (79.1%)
P322	246/407 (60.4%)
P411	334/374 (89.3%)
P423	346/420 (82.4%)
P425	261/442 (59.0%)
P431	375/431 (87.0%)
P611	267/355 (75.2%)
P613	248/429 (57.8%)
P621	361/432 (83.6%)

TABLE II
CLASSIFICATION RESULTS ON THE 2009 DATA

Data Index	# of correctly classified patterns
P123	484/806 (60.0%)
P131	470/754 (62.3%)
P132	272/626 (43.5%)
P212	478/700 (68.3%)
P213	341/720 (47.4%)
P221	539/755 (71.4%)
P222	565/691 (81.8%)
P322	369/593 (62.2%)
P411	504/561 (89.8%)
P423	599/672 (89.1%)
P425	329/680 (48.4%)
P431	574/652 (88.0%)
P611	406/567 (71.6%)
P613	395/662 (59.7%)
P621	545/652 (83.6%)

classifies the current year’s data. We constructed the SVM based classification system by using 2008 nursing-care data (15 kinds of data sets). Table I shows the classification results when the 10-fold cross validation (10CV) was used on the training data set. We can see that we have good results even when the 10CV was used. This means that the SVM based classification system has good classification performance on the testing data set.

On the other hand, Table II shows the classification results on the completely unseen data set (i.e., 2009 nursing-care data). From this table, we can see that there is classification performance deterioration on some data sets. Especially, P131, P213 and P425 have performance deterioration over 7%. We consider that this difference between the results on 2008 and 2009 data is caused by the following reasons.

- 1) There are some differences between TL_{08} and TL_{09} .
- 2) Expert’s decision may be wavering every year.

Question P425	
JPN:	＊＊さんの痛みの治療について医師にどのように働きかけていますか？＊＊さんにそのような状況がない場合は、違う事例や他の身体的苦痛に置き換えて考え、具体的にお答えください。
ENG:	How did you work on the doctor about the medical treatment of your patient's pain? When your patient does not have such a situation, please transpose to different example and other physical pain, think, and answer concretely.
Class 4 (very good answers)	
A1.	JPN 牽引施行にて神経症状が悪化し疼痛も出現しているため、牽引の評価・中止を医師に提案・依頼。
	ENG Since neurological symptoms got worse by traction enforcement and sharp pain has also appeared, I made a proposal and a request to my doctor, about evaluation and a stop of traction.
	Morphological analysis 牽引/施行/にて/神経症/状/が/悪化/し/疼痛/も/出現/し/て/いる/ため/、/牽引/の/評価/・/中止/を/医師/に/提案/・/依頼/。
	Feature vector (牽引,施行,神経症,状,悪化,し,疼痛,出現,いる,評価,中止,医師,提案,依頼)
	JPN
	Feature vector (Traction, enforcement, neurosis, state, getting worse, do, sharp pain, appear, is, evaluation, stop, doctor, proposal, request)
	ENG
Class 3 (good answers)	
A2.	JPN: 現在の内服では疼痛は軽減しない。症状が増悪しているのではないかと。薬剤の増量か、薬剤の変更を医師に依頼する。
	ENG: Sharp pain is not reduced in the present internal use. Is not the condition getting worse? I request my doctor to increase in quantity of a medicine or change a medicine.
Class 2 (almost good answers)	
A3.	JPN: 現在の鎮痛剤では、解熱せずに苦痛が軽減できていない。
	ENG: In the present painkiller, fever is not alleviated and pain cannot be reduced.
Class 1 (bad answers)	
A4.	JPN: 今の鎮静剤でコントロールできているため医師に働きかけたことはない。
	ENG: I have not worked on a doctor, since it is controllable with the present sedative.
A5.	JPN: 特に行っていない。
	ENG: I have not worked on a doctor.

Fig. 2. Examples of nursing-care texts and morphological analysis

TABLE III
DIFFERENCE BETWEEN TL_{08} AND TL_{09}

Data Index	# of terms appeared in both TL_{08} & TL_{09}	# of terms appeared in only TL_{09}
P131	489	405
P213	817	622
P425	583	599

We know the second reason is a very important problem which should be tackled. However let us assume that there is no wavering on expert's decision. In this paper, we consider only the first reason described above. We aim to reduce the differences between TL_{08} and TL_{09} , as much as possible.

Table III shows statistical information on the differences between TL_{08} and TL_{09} on P131, P213 and P425. From this table, we can see that almost 500 to 800 terms appeared in both TL . 400 to 600 terms appeared only in the TL_{09} . The features corresponding to these 400 to 600 terms in the 2009 nursing-care texts have no influence on classification, because such features were not used for training SVM based classification system.

Fig.3 shows classification results for the number of unknown terms which nursing-care texts contain. From this figure, it is shown that there is performance deterioration when the texts have two or more unknown terms. In Fig.3, classification performance of the case which has no unknown

term is deteriorated from results of 10CV on the training data set. This means that the nursing-care text data set has some nature such as wavering on expert's decision.

IV. USING ADDITIONAL TERM INFORMATION FROM WEB

A simple technique is to use known terms in the term list which are most similar to the unknown terms. In order to consider that the unknown terms in the new data set (i.e., the evaluation data set which is different from the training data set) are approximately the same as the known terms, information from Web is used. We also have been considering using conceptual sets of terms. In this paper, we describe a technique of using information from Web.

A. Unknown terms

Here we define again the unknown terms as (3)-(5). The "unknown terms" are not stored terms in the term list.

$$KT_{last} = \{kt_i | kt_i \in TL_{last}, i = 1, 2, \dots, N_{kt}\}, \quad (3)$$

$$UT_{current} = \{ut_j | ut_j \notin TL_{last}, ut_j \in TL_{current}, j = 1, 2, \dots, N_{ut}\}, \quad (4)$$

$$TL_{current} = KT_{last} + UT_{current}, \quad (5)$$

where kt_i is a known term, ut_j is an unknown term, KT_{last} is a known term set at the last year, and $UT_{current}$ is an unknown term set at the current year.

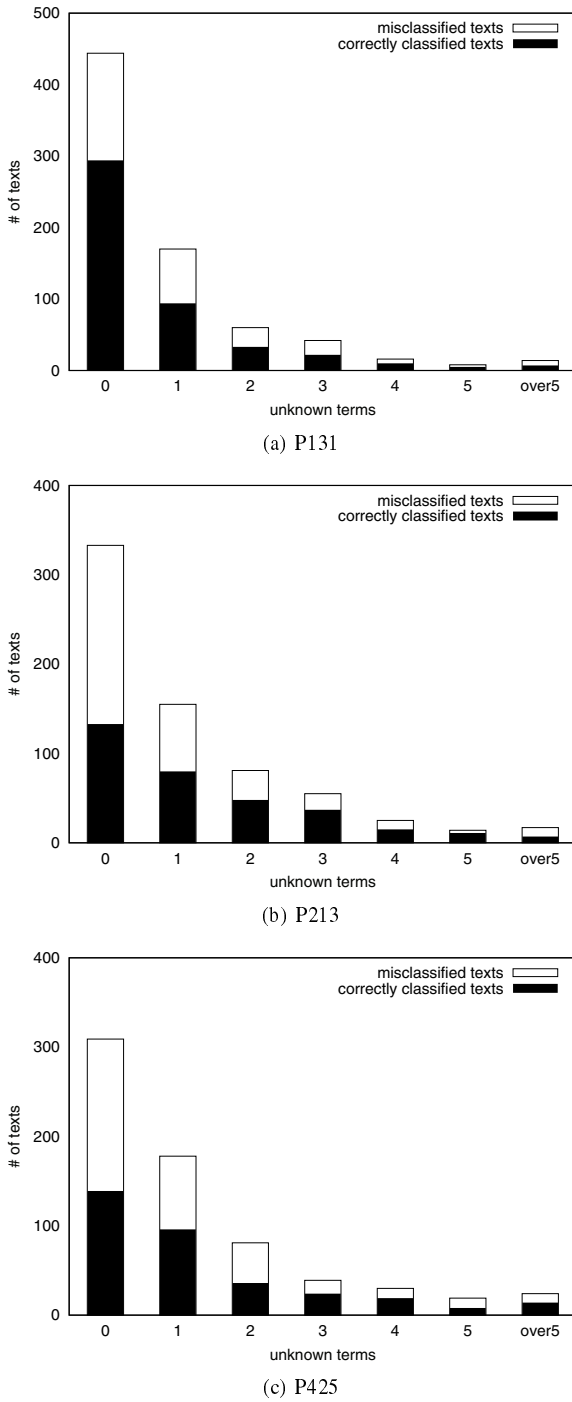


Fig. 3. Classification results for the number of unknown terms

B. Retrieving information from Web

When we use these unknown terms for classification the nursing-care texts, the information of similarity between unknown terms and known terms is needed. However, because there are only the nursing-care text sets of the last and current year, we have no additional information on the terms. In order to obtain information on the unknown terms, we try to retrieve information from the Internet. That is, each unknown term ut_j

is used as a query to a search engine.

P best results (i.e., P pages of the search results) are used. The morphological analysis is applied to the texts in the `<body>` tag of the top P results. We extract nouns from the results of the morphological analysis. A set of nouns NS_{ut_j} for the unknown term ut_j is defined as,

$$NS_{ut_j} = \{noun_k | noun_k \in SR, k = 1, 2, \dots, N_{noun}\}, \quad (6)$$

where SR is a noun set which is extracted from the set of search results. For selecting terms related to the unknown term ut_j , we consider the following three types of methods.

1) *Frequency of terms*: Then we count frequency of each noun in the search results. N_{FT} best nouns which are frequent terms are selected as the related terms.

$$FT_{ut_j} = nbest_n(\argmax_{noun_k \in NS_{ut_j}} freq(noun_k)), \quad (7)$$

where $nbest_n()$ is a function which returns a subset of n -best solutions. In this paper, we use $n = 50$. On the known terms in the TL_{last} , we also apply the above mentioned procedure.

2) *Weighted frequency by pages*: When a $noun_k$ appeared several times in a single result, we consider the $noun_k$ is used in the specific situation. On the other hand, when a $noun_k$ appeared in some results, we consider the $noun_k$ has much relevance to the unknown term ut_j . (7) is modified by using the number of pages S in which the $noun_k$ appeared, as follows:

$$FT_{ut_j} = nbest_n(\argmax_{noun_k \in NS_{ut_j}} freq(noun_k) \cdot S/N_{SP}), \quad (8)$$

where N_{SP} is the number of selected pages from search results. In this paper, we use $N_{SP} = 10$.

3) *Weighted frequency by the distance*: When a $noun_k$ appeared near the unknown term ut_j , we consider the $noun_k$ has some importance to represent the ut_j . Therefore, using the inverse of the distance between the ut_j and the $noun_k$, we modify (7) as follows:

$$dfreq(noun_k) = \sum_{p=1}^P \sum_{noun_k \in p} 1/D(ut_j, noun_k), \quad (9)$$

$$FT_{ut_j} = nbest_n(\argmax_{noun_k \in NS_{ut_j}} dfreq(noun_k)), \quad (10)$$

where $D(ut_j, noun_k)$ is the distance between the ut_j and the $noun_k$, P is the number of pages of search results.

C. Selecting the similar term

Similarity between an unknown term and the known terms is measured using the cosine similarity $sim()$.

$$st_{ut_j} = \argmax_{kt_l} sim(FT_{ut_j}, FT_{kt_l}). \quad (11)$$

A known term which has the highest similarity value to the unknown term is selected as the similar term of the unknown term.

TABLE IV
CLASSIFICATION RESULTS ON THE 2009 DATA BY USING PROPOSED THREE WEIGHTING SCHEMES

Data Index	Frequency	Weighted Frequency by pages	Weighted Frequency by the distance
P123	485/806 (60.2%)	484/806 (60.0%)	484/806 (60.0%)
P131	474/754 (62.9%)	483/754 (64.1%)	473/754 (62.7%)
P132	272/626 (43.5%)	268/626 (42.8%)	278/626 (44.4%)
P212	474/700 (67.7%)	478/700 (68.3%)	477/700 (68.1%)
P213	341/720 (47.4%)	348/720 (48.3%)	344/720 (47.8%)
P221	538/755 (71.3%)	539/755 (71.4%)	538/755 (71.3%)
P222	563/691 (81.5%)	564/691 (81.6%)	564/691 (81.6%)
P322	369/593 (62.2%)	373/593 (62.9%)	364/593 (61.4%)
P411	502/561 (89.5%)	503/561 (89.7%)	504/561 (89.8%)
P423	599/672 (89.1%)	599/672 (89.1%)	599/672 (89.1%)
P425	330/680 (48.5%)	332/680 (48.8%)	339/680 (49.9%)
P431	574/652 (88.0%)	574/652 (88.0%)	574/652 (88.0%)
P611	408/567 (72.0%)	408/567 (72.0%)	408/567 (72.0%)
P613	383/662 (57.9%)	384/662 (58.0%)	389/662 (58.8%)
P621	545/652 (83.6%)	545/652 (83.6%)	545/652 (83.6%)

V. EXPERIMENTAL RESULTS

In this section, computational experiments are described. 15 kinds of text sets collected in 2008 and 2009 were used for computer simulations. The 2008 nursing-care text sets were used for training. And the 2009 nursing-care text sets were used for testing. In the experiments, when a similar term selected by our proposed method has a similarity below the predefined arbitrary threshold value, we discarded it. It is because we consider that the terms with small similarity are not similar to the known terms.

Table IV shows the classification results on completely unseen data sets (i.e., 2009 nursing-care texts) when unknown terms are replaced by our proposed method. From experimental results, the classification performance was improved a little by our proposed method. Especially, the weighting schemes were effective over almost all data sets used in the experiments. However, since we have a slight performance improvement of classification, we have to advance our research.

VI. CONCLUSION

In this paper, to improve performance of the nursing-care text classification, a mechanism of retrieving terms from Web was proposed. The training data sets for constructing SVM based classification system are different from the evaluation data sets. That is, there are differences between the term lists of the nursing-care texts because the nursing-care texts are collected and evaluated every year. To cover the difference between the term lists of the nursing-care texts, a new term which appeared only in the evaluation data sets was used as a query of a search engine. The terms in the term list were also used as queries. Terms are represented by the search results from Web, and then are compared with each other. We used the most similar term in the term list as an alternative of the new term. From experimental results, the effectiveness of our proposed method was shown.

REFERENCES

- [1] T. Sakaguchi, M. Nii, Y. Takahashi, A. Uchinuno, R. Sakashita, "Nursing-care Data Classification using Fuzzy Systems", Proc. of SCIS & ISIS 2006, Tokyo, Japan, pp. 2017–2022, 2006.
- [2] M. Nii, Y. Takahashi, A. Uchinuno, R. Sakashita, "Nursing-care Data Classification Using Neural Networks", Proc. of 2007 IEEE ICME International Conference on Complex Medical Engineering, Beijing, China, pp. 431–436, 2007.
- [3] M. Nii, S. Ando, Y. Takahashi, A. Uchinuno, R. Sakashita, "Nursing-care Freestyle Texts Classification using Support Vector Machines", 2007 IEEE International Conference on Granular Computing, CA, pp. 665–668, 2007.
- [4] M. Nii, S. Ando, Y. Takahashi, A. Uchinuno, R. Sakashita, "Feature extraction from nursing-care texts for classification", Proc. of 6th International Forum on Multimedia and Image Processing, in CDROM (6pages), 2008.
- [5] M. Nii, S. Ando, Y. Takahashi, A. Uchinuno, and R. Sakashita, "GA based Feature Selection for Nursing-Care Freestyle Text Classification", Proc. of Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on advanced Intelligent Systems, pp. 756–761, 2008.
- [6] M. Nii, T. Yamaguchi, Y. Takahashi, R. Sakashita and A. Uchinuno, "Analysis of Nursing-care Freestyle Japanese Text Classification Using GA-based Term Selection", Proc. of World Automation Congress 2010, Kobe, Japan, IFMIP495 (CD-ROM, 6 pages), 2010.
- [7] M. Nii, T. Yamaguchi, Y. Mori, Y. Takahashi, R. Sakashita, and A. Uchinuno, "Classification of Nursing-care Data using Additional Term Information", Proc. of SCIS & ISIS 2010, Okayama, Japan, pp. 1469–1474, 2010.
- [8] M. Nii, T. Yamaguchi, Y. Takahashi, A. Uchinuno, R. Sakashita, "Improving Classification Performance of Nursing-Care Text Classification System by Using GA-based Term Selection", Journal of Advanced Computational Intelligence and Intelligent Information, Vol. 14, No. 2, pp. 142–149, 2010.
- [9] MeCab, Yet Another Part-of-Speech and Morphological Analyzer, <http://mecab.sourceforge.jp/>.