Automatic Classification and Tagging of Blog Articles with the Aim of Supporting Caregivers of Depressed Family Members

Toshihide Saito[1], Eiji Aramaki[2], Mai Miyabe[3], and Keiji Hirata[1]

***Abstract*—People who are close to and help depressed patients, as well as peripheral nurses, could be supported by performing information retrieval to eliminate or mitigate stress. It is difficult to find the information using existing keyword searches because the information caregivers need is specifically that about others in a situation similar to caregivers. To solve this problem, we introduce the structure of viewpoint, situation, and intention and take into account the context in which the structure occurs. The structure is represented using the tags generated by Conditional Random Fields (CRF). The context is represented using the labels recognized by Support Vector Machine (SVM). By 10-fold cross-validation of learning data, we evaluated a prototype system and have obtained a precision rate of 54.2 % for the context classification by SVN and 45.3% for the tag assignment by CRF. In the future, it is necessary to improve the accuracy of the context classification and tag assignment to achieve a dedicated search.**

***Keywords*—depression, classification, family caregiver, tagging, SVM, CRF.**

I. INTRODUCTION

DEPRESSION interferes with daily life and causes pain for both the patient and those who care about him or her [16]. While opportunities for a patient to get useful information have been increasing recently, useful information for caregivers is still less readily. Besides this caregivers are not focused on as much as patients in medical practice and are rarely a target for research. Many researches have been improving the medical services [2] [6] and supporting the activities involved in caregivers' nursing [8] [11]. Yamashita et all interviewed 15 caregivers who have experiences of supporting depressed family member and suggested that the use of information technology may improve the communication environment and reduce social stress [12]. Furthermore, they also claimed the technologies should contribute to *Support for Adaptation to Changes*, *Places to Share*, and *Place to Learn*. The purpose of this study is to support the information retrieval in *Place to Learn* for caregivers. Here, a caregiver is the person who cares about a patient, for example, a family member, lover, friend, or colleague. Caregivers would not only be able to obtain knowledge about their situation, but also learn about the experiences of other caregivers who have been placed in a similar situation. Such information is useful when facing troubles or predicting how the illness might affect them in the near future. In addition, sharing the information of people who faced or are now facing similar situations may reduce the stress level of caregivers. Therefore, a method is required to seek people placed in a situation close to that of each caregiver.

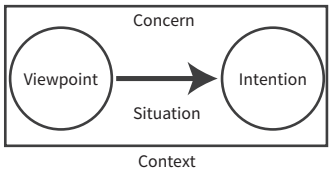

Figure. 1. Configuration of concern required by caregiver

A keyword search on the World Wide Web is a general way to look for information. However, in the case of caregivers, it often produces many irrelevant results because the information that they need is not superficial but semantically related to the troubles that they meet. For instance, the troubles are related to the relationship with a patient and other persons, family structure, annual income, and someone's actions and remarks. It has been found that people frequently read blog articles and messages on online forums about caregivers' concerns and situations, although many of the articles and messages are posted by anonymous users. Unfortunately, it is generally difficult for a computer to look for such information by matching the keywords because such articles and messages are not written in a machine-friendly style, nor do they follow a specific document configuration. Therefore, we think the conventional techniques of information retrieval are unsuitable for our purpose. Our research aims at developing a technique for semantic information retrieval from the non-uniform sentences of online forums or blogs. In particular, we aim to achieve the retrieval of similar documents required by caregivers using the context of concerns and the structure of viewpoint, situation, and intention (Figure.1).

II. CLASSIFICATION AND TAGGING WITH MACHINE LEARNING

Various techniques for information searching have been developed, for example sentiment features [10], and the cloud services meeting users' needs [15]. However, to get better information for caregivers, more detailed information is needed. There is a Semantic Textual Similarity (STS) task, used as a method for retrieval of similar documents [4]. The major aim of STS is to quantify the similarity of a pair of two short sentences. Severyn used syntactic structures for learning [1]. Beltagy used high-speed logical reasoning

[1] T. Saito and K. Hirata are Future University Hakodate, Japan. (e-mail: g2114012@fun.ac.jp, hirata@fun.ac.jp)
[2] E. Aramaki is Nara Advanced Institute of Science and Technology, Japan (e-mail: aramaki@is.naist.jp).
[3] M. Miyabe is Wakayama University, Japan (e-mail: miyabe@sys.wakayama-u.ac.jp).

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

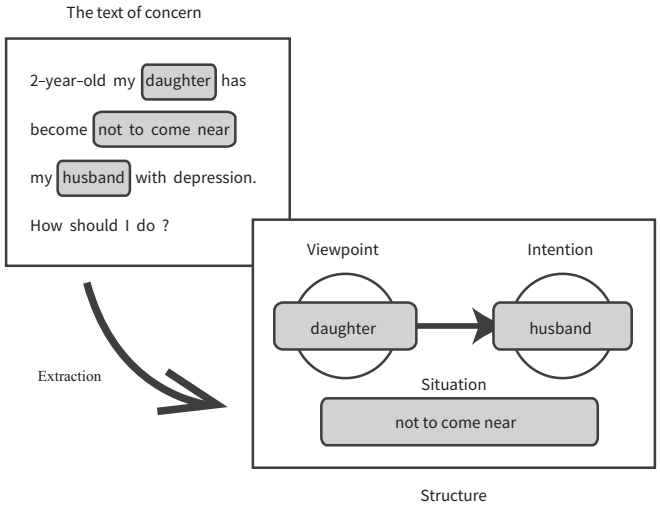

Figure. 2. Extracting the structure of text of concern

with probability [7]. These approaches are effective and the STS task targets only two sentences, but we have to deal with two or more sentences in documents. Thus, a possible approach is to introduce semantic items. Semantic items should be represented in such a way that a computer can recognize the concerns caregivers have. For example, these items include standpoint, time, relationship, and background. However, to list and represent each item is not effective because there are too many possible combinations of items and a huge amount of data would be needed to realize machine learning that could process them. To realize the retrieval of similar documents, we divide the process into two stages: classification of the context, and tagging for extraction of the structure. In classification of the context, a document is labeled according to the relationship between the author and the patient: spouse (wife or husband), sibling, parents, and others. In tagging for extraction of the structure, a document is structurized by inserting tags into it; the tags represent the detailed information that is different from the relationship between an author and a patient: viewpoint, situation, and intention. In Figure. 2, "daughter" is the viewpoint, "husband" the intention, and "not to come near" the situation.

To implement our method, we employ Support Vector Machine (SVM) which is the typical method for solving the classification problem. SVM is used in the task of Recognizing Textual Entailment (RTE) which recognizes the semantic connection of two sentences [14]. Next, we employ Conditional Random Fields (CRF) [9], which is often used for solving the sequence labeling problem. CRF is effective in data mining to extract information [3]. As stated above, our retrieval method consists of two stages: SVM recognizes the relationship between an author and a patient, and CRF structurizes a document in terms of viewpoint, situation, and intention.

A. Context in Retrieving Relevant Blog Articles

There are cases in which fragments of texts are identical but have different meanings. For example, the information contained in a forum or blog entry is different depending on whether the patient is a husband or a wife. It is necessary to represent and match the context of the structure, for which SVM is employed. The labels used in SVM classification are set up by a manual pre-classification of 100 documents (Table I). Because many documents belong to the label of "the patient is a husband", we elaborated on this by splitting the label into the labels of "the patient is a husband with no children or unknown" (X1) and "the patient is a husband with a child/children" (X2). On the other hand, the labels of "the patient is a parent" and "the patient is a parent of spouse" are not frequently used. So, we have merged them, producing the label "the patient is a parent (including the parent of spouse)" (X4). In the same manner, finally we have made the label of "the patient is a brother or sister" (X5).

Table I
TYPES OF LABELS

Label name	Who is a patient
X1	Husband with no children or unknown.
X2	Husband with a child/children
X3	Wife
X4	Parent
X5	Brother or Sister
X6	Other

Table II
TYPES OF TAGS

Tag name	Whose behavior, action or state
a	Author
p	Patient
o	Other
Tag name	**From whom to whom**
ap	From author to patient
pa	From patient to author
do	From author or patient to other
od	From other to author or patient
oo	From other to other

B. Structure in Retrieving Relevant Blog Articles

The structure of viewpoint, situation, and intention is represented by tags (Figure. 3); *a*, *p*, and *o* are used as the basic tags. Each of the tags means "author", "patient", and "other", respectively, and by combining them, we create tags which have direction (Table II). We merge *po* and *ao* to create

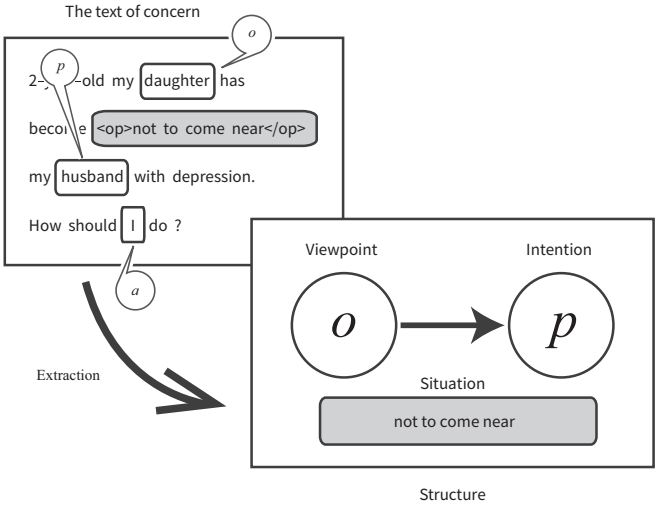

Figure. 3. The example of tagging for extraction of the structure

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Table III
THE RESULT OF 10-FOLD CROSS-VALIDATION OF LEARNING DATA OF CLASSIFICATION BY SVM

the number of matching data / the number of correct data

	A	B	C	D	E	F	G	H	I	J	AVE
X1	4/9	5/8	4/8	5/8	2/8	5/8	1/8	4/8	5/8	4/8	3.9/8.2
X2	5/7	5/7	2/7	3/7	2/7	4/7	1/7	4/7	3/8	4/8	3.3/7.4
X3	7/8	4/8	6/8	5/8	4/7	6/7	4/7	4/7	4/7	6/7	5.0/7.7
X4	8/10	9/10	8/10	9/11	8/11	9/11	8/11	8/11	10/11	8/11	8.5/11.1
X5	2/7	3/7	3/7	3/7	2/7	4/7	3/7	2/7	5/7	1/7	2.8/7.5
X6	0/5	2/5	2/5	1/5	1/5	1/5	2/5	0/5	1/5	1/5	1.1/5.6
SUM	26/46	28/45	25/45	26/46	19/45	29/45	19/45	22/45	28/46	24/46	24.6/45.4

Precision rate 54.2% [= 24.6 / 45.4)

Table IV
THE RESULT OF 10-FOLD CROSS-VALIDATION OF LEARNING DATA OF TAGGING BY CRF

	A	B	C	D	E	F	G	H	I	J	Average
CORRECT	1181	1501	1472	1357	1198	1389	1282	1433	1169	1374	1335.6
SYS	666	773	868	737	660	726	728	797	594	787	733.6
MATCHED	322	324	370	337	312	329	352	334	277	367	332.4

Precision rate 45.3% [= 332.4 / 733.6), Recall rate 24.9% [= 332.4 / 1335.6), F-measure 0.321

do, and *op* and *oa* to create *od*. This is because these tags are less likely to appear.

III. EXPERIMENT AND EVALUATION

A. Building Corpus

We obtained sentences for the experiment from the OKWAVE web site [13], which provides an online knowledge community in Japan. The way it works is that a user posts a question, and then some of other users who are interested may concern respond. To post, a user needs to perform free member registration, but uploaded documents are available to anyone on the Internet. We acquired 3577 documents in the married couple and family category of OKWAVE on August 26, 2014, and selected 449 documents that appeared to be written by caregivers. On average, these documents consisted of 1656 characters in Japanese. We built a corpus of data by classifying and tagging them.

B. Evaluation of Classification

Firstly, we replace the morphological analysis result of documents with binary vectors. Secondly, we input it to TinySVM [17] which is set as a polynomial function of the 2nd degree. TinySVM is an open source SVM software. In classification, we use one-versus-rest method. In evaluation, we use 10-fold cross-validation. Table III shows the results. The 10-fold divided data are titled with letters from A to J. The recall rate is always at maximum value, because all documents are always classified to one of the labels in Table I. Therefore, we show only the precision rate for evaluation of classification. The highest value is obtained for X4 "the patient is a parent (including the parent of spouse)". The lowest value is for X5 "the patient is a brother or sister", excluding X6 "the patient is other". However, the accuracy of each label is better than the probability of the chance level (1.67 = 10 / 6).

C. Evaluation of Tagging

We give the morphological analysis result of documents as a sequence and labeled sequence of IOB2 tag [5] to CRF++ [18] which is a CRF tool kit. In evaluation, we again use 10-fold cross-validation. We present precision rate, recall rate, and F-measure for evaluation. For example two sentences of "at night do shopping" in Figure. 4 can potentially be tagged in two ways, although the meanings of sentences are the same. Therefore, we cannot determine whether the result of tagging is correct or not, from only the start and end points of the tags. Thus, we adopt the tagging labels in the IOB2 format.

Table IV shows the number of the tags assigned. The 10-fold divided data are titled with letters from A to J. The precision rate is 45.3% [= 332.4 / 733.6), the recall rate is 24.9% [= 332.4 / 1335.6), and F-measure is 0.321. The recall rate is lower than the precision rate, which means that many of the features to be tagged were missed. Also the precision rate is lower than 50% unfortunately.

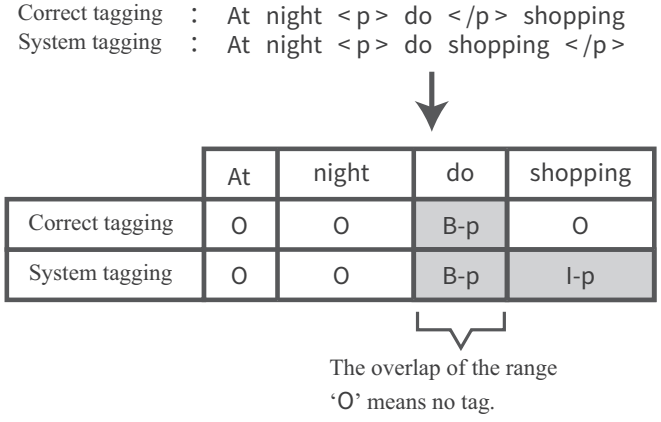

Figure. 4. The example of tagging in two ways

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

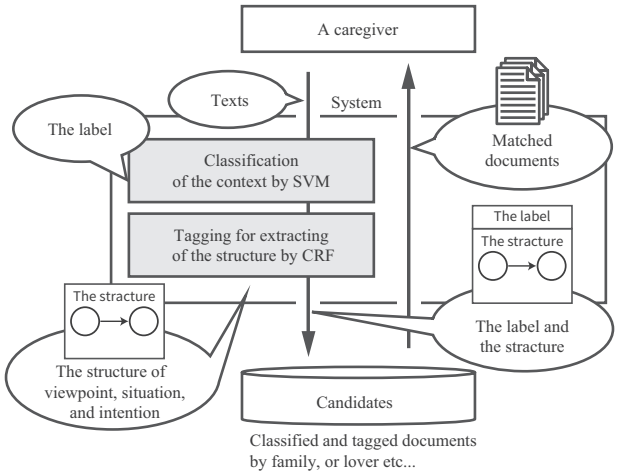

Figure. 5. System diagram

D. DISCUSSION

In evaluation of classification, the accuracy of each label is better than the chance level probability. This indicates that the classification of the context by SVM is effective for the documents written by caregivers. The highest value is for X4, "the patient is parent (including a parent of spouse)", in which "the patient is a parent" and "the patient is the parent of spouse" have been combined. This is a case in which merging is effective, because "the patient is a parent" is similar to "the patient is the parent of spouse" at the notational level. On the other hand, the value of X5, "the patient is a brother or sister", which combines "the patient is an older brother or sister" and "the patient is a younger brother or sister", is not as high as the value of X4. This is probably because they are not similar at the notational level. It leads us to conclude perform a similarity search, it is necessary to consider the similarity of each label at the notational level.

Evaluation of tagging reveals that the recall rate should be improved. Regarding the precision rate, there are two possible reasons why it is lower. One is mistaking a tag name, and the other is that many of the features to be tagged were missed. In general, the longer the range of tagging, the worse the accuracy of tagging. For that reason, it is possible to improve F-measure by targeting only single nouns. This means the extraction of the information for caregivers by CRF could be more effective.

IV. CONCLUSION

In this paper, we have presented an information search method for caregivers of a depressed family members, and the results of 10-fold cross-validation of the learning data of SVM and CRF. The accuracy of classification by SVM is more than 50% as a whole, and the accuracy of label identification is more than 40%.

We have examined the method which takes into account the extent to which the ranges tagged by CRF overlap with that of sentences in the corpus. The value of F-measure is 0.321, which is inadequate for practical use. We expect that the value of F-measure could be improved by targeting only nouns.

Future work includes building a working information retrieval system for family caregivers, the system diagram of which is shown in Figure. 5. Firstly, a caregiver, as a user, inputs a text of some sentences about their concerns into the system. Secondly, the system extracts the context and the structure from the text. Finally, the system outputs similar documents from tagged and classified candidates for the caregiver.

V. ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 70396141.

REFERENCES

[1] A. Severyn, A. Moschitti, and M. Nicosia, "Learning semantic textual similarity with structural representations," *The 51st Annu. Meeting of the Association for Computational Linguistics* (*ACL 2013*), Sofia, Bulgaria, 2013, pp. 714-718.

[2] D. Coyle, G. Doherty, M. Matthews, and J. Sharry, "Computers in talk-based mental health interventions," in *Interacting with Computers*, 2007, vol. 19, Issue. 4, pp. 545-562.

[3] D. Pinto, A. McCallum, X. Wei, and W. B. Croft, "Table Extraction using Conditional Random Fields," *Proc. 26th Annu. int. ACM SIGIR Conf. Research and development in information retrieval* (*SIGIR '03*), Toronto, Canada, 2003, pp. 235-242.

[4] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, "SemEval-2012 task 6: A pilot on semantic textual similarity," *1st Joint Conf. Lexical and Computational Semantics* (**SEM*), Montreal, Canada, 2012, pp. 385-393.

[5] E. F. Tjong Kim Sang and J. Veenstra, "Representing Text Chunks," *Proc. 9th Conf. European chapter of the Association for Computational Linguistics* (*EACL 1999*), Bergen, Norway, 1999, pp.173-179.

[6] G. Doherty, D. Coyle, and J. Sharry, "Engagement with Online Mental Health Interventions: An Exploratory Clinical Study of a Treatment for Depression," *The ACM SIGCHI Conf. Human Factors in Computing Systems* (*CHI 2012*), Austin, USA, 2012, pp. 1421-1430.

[7] I. Beltagy, K. Erk, and R. Mooney, "Probabilistic soft logic for semantic textual similarity," *The 52st Annu. Meeting of the Association for Computational Linguistics* (*ACL 2014*), Baltimore, USA, 2014, pp. 1210-1219.

[8] J. Duncan, L. J. Camp, and W. R. Hazelwood, "The portal monitor: A privacy-enhanced event-driven system for elder care," *Proc. 4th Int. Conf. Persuasive Technology* (*Persuasive 2009*), Claremont, USA, 2009, pp. 1-9.

[9] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," *Proc. 18th Int. Conf. Machine Learning* (*ICML 2001*), Williamstown, USA, 2001, pp. 282-289.

[10] K. Minami et al., "Comprehensive Web Search based on Sentiment Features," *Proc. Int. MultiConference of Engineers and Computer Scientists 2014* (*IMECS 2014*), Hong Kong, China, 2014, pp. 483-488.

[11] L. S. Liu et al., "Improving communication and social support for caregivers of high risk infants through mobile technologies," *The 2011 ACM Conf. Computer Supported Cooperative Work* (*CSCW*), Hangzhou, China, 2011, pp. 475-484.

[12] N. Yamashita, H. Kuzuoka, K. Hirata, and T. Kudo, "Understanding the conflicting demands of family caregivers caring for depressed family members," *Proc. SIGCHI Conf. Human Factors in Computing Systems* (*CHI 2013*), Paris, France, 2013, pp. 2637-2646.

[13] OKWAVE. *OKWAVE*. http://okwave.jp, accessed on Jan, 12, 2016.

[14] P. Malakasiotis and I. Androutsopoulos, "Learning Textual Entailment using SVMs and String Similarity Measures," *Proc. ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, Czech Republic, 2007, pp. 42-47.

[15] S. Gong and K. M. Sim, "CB-Cloudle: A Centroid-based Cloud Service Search Engine", *Proc. Int. MultiConference of Engineers and Computer Scientists 2014* (*IMECS 2014*), Hong Kong, China, 2014, pp. 446-451.

[16] The National Institute of Mental Health. *The National Institute of Mental Health: NIMH Depression.* https://www.nimh.nih.gov/health/topics/depression/index.shtml, accessed on Jan, 12, 2016.

[17] T. Kudo. *TinySVM: Support Vector Machines - ChaSen.org*. http://chasen.org/~taku/software/TinySVM/, accessed on Jan, 12, 2016.

[18] T. Kudo. *CRF++: Yet Another CRF toolkit.* https://taku910.github.io/crfpp/, accessed on Jan, 12, 2016.

ISBN: 978-988-19253-8-1
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)