Performance Improvement of Drug Effects Extraction System from Japanese Blogs

Shiho Kitajima, Rafal Rzepka, Kenji Araki
Graduate School of Information Science and Technology, Hokkaido University
shihov_vo@media.eng.hokudai.ac.jp

Abstract—**Information disclosed to the public by patients is very important for people who are suffering from same illness because such information can be a source of knowledge and encouragement. Our aim is to make a system that extracts, organizes and visually represents information from patients' blogs. As the first step, the purpose of this paper is to extract descriptions of the effects caused by taking drugs as a triplet of expressions - drug name, object of change, and its effect - from illness survival blogs. However, conventional extraction methods are not suitable since these blogs are written in free natural language. Therefore, this paper proposes a method to extract the triplets using specific clue words and parsing the results. An evaluation experiment confirmed that medication usage information can be extracted with high accuracy using our proposed method, in comparison to existing methods. Moreover, recall was improved by combining our proposed method and a baseline system.**

Keywords-component; Text mining, Opinion mining, Information extraction, Medication usage information

I. Introduction

With the rapid growth of the Internet, it is becoming easier for people to write their opinions, behavior, and interests on the Web in real time. Japanese people usually disclose personal information anonymously on social media and blogs. It is considered that anonymous narratives are unreliable due to lacking a sense of responsibility. However, anonymized information on the Web gives us information that may be inaccessible in the real world, since anonymity allows users to express opinions and ideas that they may not want someone close to know, without the need to worry about appearances. Additionally, information on the Web is extensive and instantaneous, and thus can be used to understand people's changes and trends. Many patients and their family members are posting information about diseases and treatments on the Web too. Information from people who have experienced illness is important because it is different from the information distributed officially or by doctors. By examining information provided by people who are in the same situation, other patients can decide to how to confront their disease, and treatment policies can be determined. This kind of health information enables patients to play an active role in their healthcare management.

To collect such information, we focused on blogs written by patients in natural language. Daily records written in blogs are very valuable because of the description of how to deal with the disease. Furthermore, they are first-hand information that cannot be obtained from textbooks written by people who have never experienced the relevant disease. However, it is not easy to extract and use this information appropriately since there is a huge number of blogs on the Web, and they are written in free natural language.

Accordingly, we aim to make a system that extracts medical information from patients' blogs, objectively organizes the information in chronological order, and visually represents it. As the first step of this task, this paper presents an approach which enables the extraction of the effects and changes caused by taking drugs as a triplet of expressions - drug name, object of change, and its effect - from illness survival blogs. Future research will analyze the polarity of this information, and judge whether the triplets of medical information extracted from patients' blogs are descriptions of changes that are desired or not. For this purpose, in this paper, we extract information for which polarity can be read not by single words (e.g. "effective", "good", "bad"), but by sets of object of change and relevant effect (e.g. "pain/sensation - loss", "hair - loss").

II. Related work

In recent years, research is being conducted on the medical applications of information technology, with a focus on natural language processing techniques[1]. Aramaki et al.[2] developed a system that examines adverse effects and event information buried in Electronic Health Records (EHR) in hospitals. As in Aramaki et al.'s research, the relationship between drugs and side effects is usually extracted by SVM[3]. However its accuracy as an extraction method is low. Shinohara et al.[4] proposed a method that extracts side effects of drugs using syntactic patterns without using SVM. In electronic medical records, expressions are recorded in written (formal) language. In blog articles, on the other hand, expressions are often written in spoken (casual) language and non-technical terms. We believe that rare information can be collected by extracting medicinal effects and side effects from texts written not from the doctor's point of view, but from the viewpoint of patients. Moreover, we aim to design a system that helps patients to support each other by using texts written from the view of patients.

III. Methodology

This paper presents an approach which uses clue words to enable the extraction of the effects and changes caused by taking drugs as triplets (“drug”, “object”, “effect”) from illness survival blogs. Here, "drug" is the name of the drug used,

978-0-7695-5119-7/19 $26.00 © 5119 IEEE
DOI 10.1109/ICSC.2013.71

IEEE computer society

"object" is an area or disposition in which the change or effect of the drug occurred, and "effect" is an expression or mental attitude that shows the effect or change. We extract from snippets, which are summary statements of blogs. More than 40,000 blog sites written by patients and their families are registered on TOBYO[2]. The entries are already tagged with the name of the patient's disease, the patient's gender and so on, so we plan to use the information in the future. According to our preliminary survey, we found that 78.3% of the information on effects of drugs was described in the same sentence that contains the name of the drug. Accordingly, in this paper, we aim to extract information on effects of drugs from sentences in snippets that contain the name of the drug. Figure 1 shows an overview of our system. A detailed explanation of the extraction process is discussed later in this section.

A. Clue words

In order to identify the position of the word that contains information on drug effects, we set specific clue words. We checked 181 sentences involving the triplet ("drug", "object", "effect") of the information on drug effects in advance. In 138 sentences (79.2% of total), we found that the object in which change and effect occurred and the effect are written after using specific expressions (underlined in Figure 1) which suggest the taking and using of drugs. We collected the expressions, and set 32 nouns and verbs (in the base form) as clue words, for example, *eikyou* (effect), *okage* (virtue), *sei* (reason), *tsukau* (use).

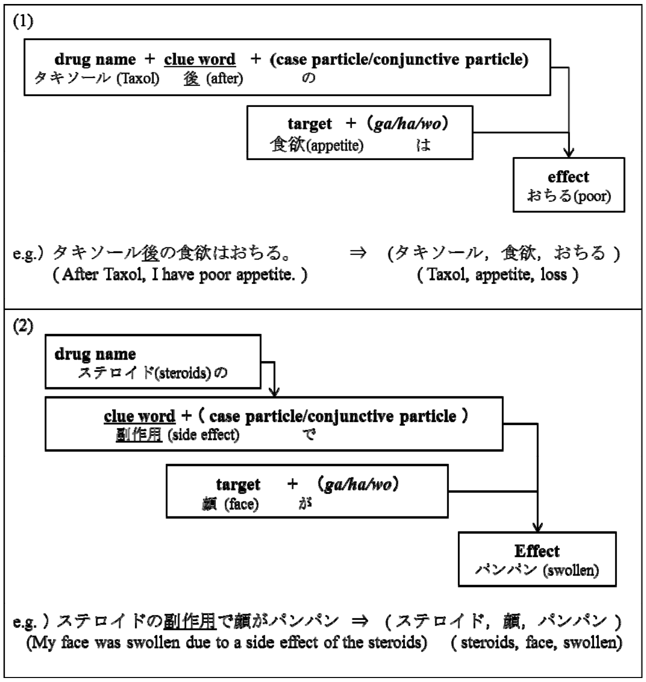
(1)
drug name + clue word + (case particle/conjunctive particle)
タキソール (Taxol) 後 (after) の
target + (ga/ha/wo)
食欲(appetite) は
effect
おちる(poor)
e.g.) タキソール後の食欲はおちる。 ⇒ (タキソール，食欲，おちる)
(After Taxol, I have poor appetite.) (Taxol, appetite, loss)
(2)
drug name
ステロイド(steroids)の
clue word + (case particle/conjunctive particle)
副作用 (side effect) で
target + (ga/ha/wo)
顔 (face) が
Effect
パンパン (swollen)
e.g.) ステロイドの副作用で顔がパンパン ⇒ (ステロイド，顔，パンパン)
(My face was swollen due to a side effect of the steroids) (steroids, face, swollen)

Figure 1. Pattern of extraction

B. Converting the snippet

The aim of converting the snippet is to reduce parsing mistakes.Our system separates snippets into statements using a delimiter "。" ("."), and extracts from the sentence that contains the name of the drug. In addition, our system deletes unnecessary parentheses (e.g. emoticons "(^_^)") from the sentences. After that, the system replaces the drug names that are not registered in the IPA dictionary used in CaboCha[3] (a Japanese dependency parser) into "MEDICENE", in order not to split incorrectly.

C. Detecting phrase position

The system detects the position of the phrase that includes the word which is an element of the triplet relating to drug effects by extraction rules (Figure 2) that use the dependency pattern and clue words. In Figure 1, rectangles indicate a phrase and arrows show a dependency parsed by CaboCha. We found that the postpositional particles used between the phrase that contains the target word and the phrase containing the effect are "*ga*" (57.7%), "*ha*" (12.7%), and "*wo*" (11.0%). Accordingly, the system extracts the target elements from the phrase depending on the phrase containing the effect by "*ga, ha, wo*". Furthermore, the system only extracts when the postpositional particle used after a clue word is a case particle (e.g. "*de*", "*kara*", "*yori*") or a conjunctive particle (e.g. "*ga*", "*node*", "*keredomo*"), because case particles have a nuance of "continuance" or "process" of actions and states, and conjunctive particles have a nuance of "cause" or "motive" for process.

D. Converting the phrase

The final component of our system is converting the phrase which contains the elements of object and effect into an appropriate form, as a triplet of drug effect information. To begin with, the system extracts, as the element, the original form of the first word from the phrase that was judged to contain the element. Our system combines words as the element according to the situation: for example, successive nouns, a prefix and the following noun, a modifier clause which has the postpositional particle "*no*" ("of") (as head of sentence). Furthermore, our system produces a negative expression of the element when "*nai*" ("not") exists in the phrase an odd number of times. In this study, our system does not consider "…*masen*" ("not"), negative clue words (e.g. "*kiru*" ("stop"), "*genryo*" ("reduce")) or contradictory conjunctions (e.g. but, however) as negative words. We plan that further studies on this system will consider such words.

IV. EVALUATION EXPERIMENT

The purpose of our experiment was to demonstrate the extraction performance of the algorithm on blog data. The first author conducted judgments on the appropriateness of the triplets extracted by our system and baseline systems for information on drug effects. We used precision, recall, and F-measure to evaluate. These can be calculated using the following three equations.

[1] An italic letters shows a Japanese character.

[2] http://www.tobyo.jp/

[3] http://chasen.org/ taku/software/cabocha

$$\text{Precision} = \frac{correct}{correct + mistake} \quad (1)$$

$$\text{Recall} = \frac{correct}{\text{340 correct triplets by manual extraction}} \quad (2)$$

$$\text{F} - \text{measure} = \frac{precision * recall * 2}{precision + recall} \quad (3)$$

A. Data

Our system extracts from 2,369 sentences containing drug names among 2,000 snippets which were collected by "TOBYO-*jiten*" ("dictionary"). TOBYO-*jiten* is a tool in TOBYO that searches blog articles using medical keywords. The 340 correct triplets of information on effects of drugs were obtained by manual extraction.

B. Baseline systems

Shown in Figure 2 are three extraction methods used as baselines for comparison. Baseline system 1 extracts when the parsing results match the pattern "drug name (*de/no/ha*) → object" and "object (*ga/ha/wo*) → effect". In baseline systems 2 and 3, the extraction methods use words in EVALDIC_ver1.0.1[7], a general Japanese dictionary of evaluation expressions, as the evaluation factors. In baseline system 2, the target elements are extracted from the phrase depending on the phrase that contains the evaluation factor (the word in EVALDIC_ver1.0.1) by "*ga, ha, wo*", from sentences that contain the name of the drug. Moreover, baseline system 3 extracts when a dependency relation using the particles "de, no, ha" is formed between the phrase that contains the drug name and the phrase that contains the target element.

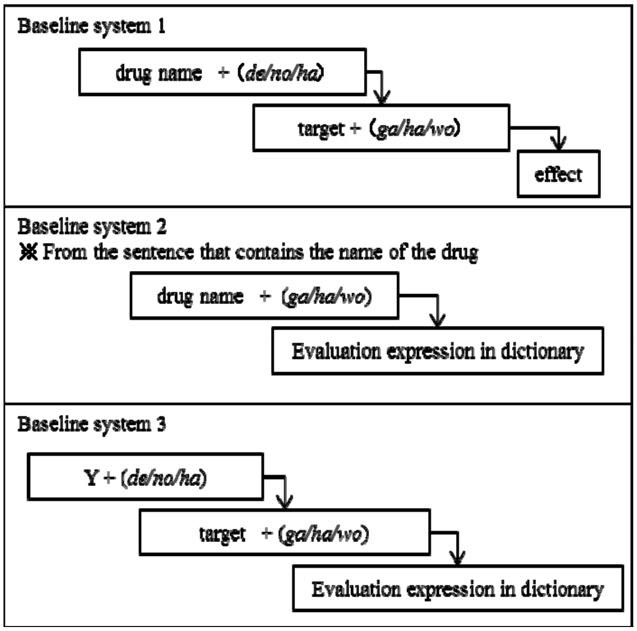

Figure 2. Baseline system extraction pattern

TABLE I. EVALUATION RESULTS

	Precision [%]	Recall [%]	F-Measure
Our system	**51.1**	7.1	12.4
Baseline system 1	10.5	3.2	5.0
Baseline system 2	11.2	28.5	16.1
Baseline system 3	8.2	1.2	2.1

TABLE II. INTEGRATED SYSTEM RESULT

	Precision [%]	Recall [%]	F-Measure
Integrated system	12.7	**33.2**	**18.4**

C. Results and discussion

The results of the evaluation are displayed in Table I. It was confirmed that an extraction system that uses clue words is effective for extracting information on drug effects from illness survival blogs. From the results of the baseline systems 2 and 3, we confirmed that the number of extractions is dependent on the limitation of the relationship between the drug name and other elements. There are two types of errors in our system: misjudgments of the position of the phrase including the element, and unsuitable conversion of the element of the triplet. The reasons why our system misjudged the position of phrases are parsing errors, which are attributed to incomplete sentences contained in snippets, and unmatched patterns of extraction due to missing postpositional particles because the blogs are written in spoken language. Moreover, some sentences could not be extracted from, owing to the use of unregistered clue words. In future studies, we will increase the number of clue words. Additionally, due to nonexistence of clue words and mismatch of the extraction patterns, some correct triplets extracted by the baseline systems 1 and 3 could not be extracted by our system. Thus, it can be considered that integration of the methods of the baseline systems into our system would be effective to increase recall.

Therefore, we designed a new system that extracts using baseline system 2 from the sentences that our system cannot extract triplets from, and performed an evaluation. Table II shows the results. The number of correct extractions increased by 89, the recall increased by 26.1 points, and the F-measure was also the highest. The number of correct triplets that could not be extracted by the combined system was only 1. This demonstrated that an integrated approach is effective in raising recall. However, the accuracy was reduced because the outputs by baseline system 2 contain many mistakes.

In order to reduce output errors, it is necessary to judge the appropriateness of the elements for the triplets of drug effect information. We often find that a triplet is wrong because the extracted object element was inadequate (e.g. error target: "*kakuritsu"* (probability), correct target: "*kami ga nukeru kakuritsu"* (probability of one's hair falling out); error target: "1*nen-ijou"* (more than 1 year), which is simply a period of time). When we describe the effects of drugs, we use various evaluation words, but the objects in which the change or effect of the drug occurred are limited to some extent. Therefore, future tasks required to solve these issues include creating a dictionary that contains words used as the object when we explain drug effects, and using machine learning to convert the phrase into an appropriate form and range, as a triplet of drug effect information.

V. Conclusion and future works

We have presented an approach to extracting information on the effects of drugs from snippets of illness survival blogs. It was confirmed that our system can extract with higher precision than existing methods. The recall increased by 26.1 points by integrating the method of baseline system 2 into our system. In future works, we will change the target data to all blog articles, increase the number of clue words in order to reduce parsing errors and match the extraction patterns, and extract with due consideration of the appropriateness of the triplet elements of drug effect information. After extraction, the next aim of our system is determining the polarity of the information and visualizing chronological changes.

References

[1] Carolin Kaiser and Freimut Bodendorf, "Mining Patient Experiences on Web 2.0 - A Case Study in the Pharmaceutical Industry.", SRII Global Conference 2012, pp.139-145, 2012.

[2] Eiji Aramaki, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, Kayo Waki, and Kazuhiko Ohe: "*Extraction of Adverse Drug Effects from Clinical Records*", Stud Health Technol Inform. 2010, pp.739-743, 2010.

[3] Vladimir Vapnik, "*Statistical Learning Theory"*, WileyInterscience, 1998.

[4] Emiko Shinohara, Keigo Hattori, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, Eiji Aramaki, and Kazuhiko Ohe, "*Koubun Patan ni Motozuku Yakuzai Fukusayou Zyouhou no Zidou Chuushutsu (Automatic Extraction of Drug Side Effects Using Syntactic Patterns)*", The 31nd Joint Conference on Medical Informatics, pp.521-524, 2011.

[5] Kenji Sugiki and Shigeki Matsubara, "*A product retrieval system robust to subjective queries*", International Journal of Product Lifecycle Management 3(2-3), pp.151-164, 2008.

[6] Masaaki Tsuchida, Hironori Mizuguchi, and Dai Kusui, "*Ranking Method of Object-Attribute-Evaluation Three-Tuples for Opinion Retrieval*", New Frontiers in Artificial Intelligence, JSAI 2008 Conference and Workshops, vol. 5447, pp.87-98, 2009.

[7] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima, "*Collecting evaluative expressions for opinion extraction*.", In Proceedings of IJCNLP, pp. 584–589, 2004.