Extraction of Drug Information Using Clue Words 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

In the medical field, daily records of the patient are very important and useful. 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 informal 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.

Key Words- Text mining, Opinion mining, Information extraction, Medication usage information

1. Introduction

With the spread of the Internet, it is becoming easier for people to write their opinions, behavior, and interests on the Web in real time. Although information on the Web is not highly reliable, it 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. In the medical field, such information from people who have experienced illness is regarded as important because it is different from the information distributed officially or by doctors. By examining such information, other patients can decide 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 informal 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").

2. 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][2][3].

Aramaki et al.[4] developed a system that examines adverse effects and event information buried in Electronic Health Records (EHR) in hospitals. The relationship between drugs and side effects is usually extracted by SVM [5][6]. Shinohara et al. [7] proposed a method that extracts side effects of drugs using syntactic patterns. 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. Because new information that is not documented in drug descriptions can be mined from blog reports and can provide valuable input to determine unknown drug effects.

Extraction of opinions about certain objects as triplets has been studied. Sugiki et al.[8] proposed a method for retrieving products that match users' search queries written in natural language using representations that

transformed texts of product reviews or queries into triplets (“object”, “item”, “value”). Tsuchida et al.[9] proposed a ranking method for opinion retrieval from weblog articles that uses a confidence model of opinion as a three-tuple of object-attribute-evaluation. The method extracts attribute-evaluation pairs (finding evaluation words using the evaluation-expression dictionary) by SVM, and associates objects with attribute-evaluation pairs by the distance of those words. In studies to extract opinions about a product or location, it is possible to extract using an evaluation-dictionary, since the object (“object of change” in our study) is clear and expressions of evaluation used in texts are limited. Blogs are written not to evaluate the effects of drugs, but to record daily life. Expressions of evaluation that are used to write about drugs or conditions contain words that are used in product reviews, as well as mimetic words. There are no evaluation-dictionaries that include Japanese mimetic words, for example "*panpan*", meaning “bulging”. Moreover, it is not easy to create an evaluation-dictionary that is suitable for medical evaluation, because of the variety of mimetic words.

3. 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, "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. We extracted 181 pieces of information on effects of drugs that are represented by triplets from snippets of blogs registered in TOBYO[1]. More than 40,000 blog sites written by patients and their families are registered on TOBYO. 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. 254 pieces of information on effects of drugs existed in random 1,000 snippets. 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.

The extraction was performed in three steps:

1) The system converts the snippet to a form that is easy to parse.
2) The system detects the position of the phrase that includes the word which is an element of the triplet relating to drug effects, using extraction rules that consider the pattern of syntactic dependency and the clue words we set uniquely.
3) The system converts the phrase into an appropriate form as an element of the triplet relating to drug effects.

Figure 1 shows an overview of our system. A detailed explanation of the extraction process is discussed later in this section.

3.1. Clue words

In order to identify the position of the word that contains information on drug effects, we set specific clue words.

A selection of sentences involving the triplet (“drug”, “object”, “effect”) of the information on drug effects is shown in Table 1. 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 Table 1) which suggest the taking and using of drugs. We collected the expressions, and set 32 nouns and verbs as clue words. The verb of the set is in the base form, in order to change surface form by conjugation. Table 2 shows a selection of clue words.

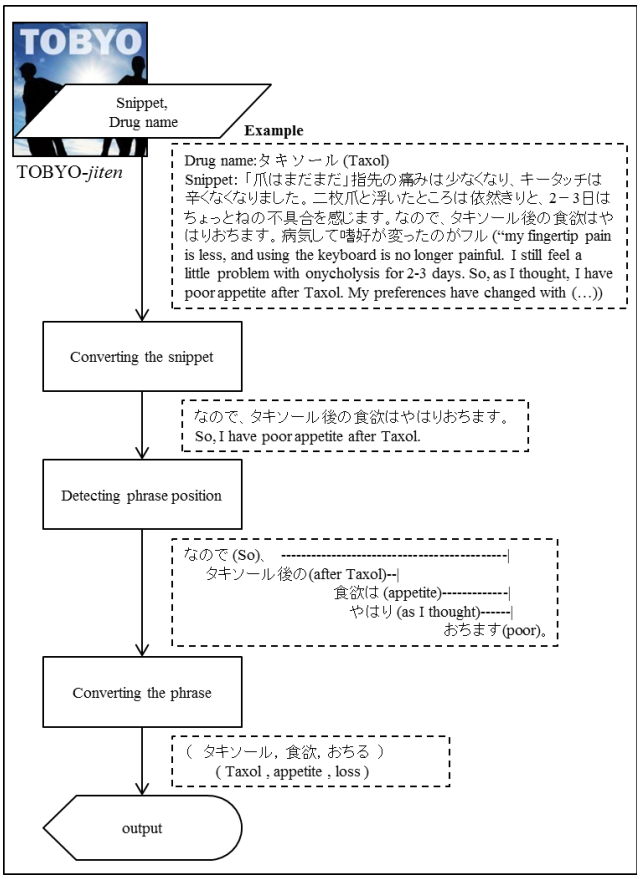

Figure 1. System overview

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

Table 1. Example sentences involving the drug effect triplet

1	それからプレドニンを飲み始めたのですが、プレドニンの副作用で骨が弱くなり (Then I started drinking prednisone, but my bones were weakened by side effects of prednisone)
2	私もおタキ様（タキソテール）の後は、肩や背中がすごく凝ります (My shoulders and back are very stiff after Mr.Taki(Taxotere) too)
3	現在の発作は、アレビアチンの注入時間前後（6時、14時、22時）に痙攣が出現している (In the current attack, convulsions appeared around the time of injecting Aleviatin (6:00, 14:00, 22:00))
4	多分ジプレキサで頭がボケ気味なんだと思います (Maybe, I am going senile due to Zyprexa)

Table 2. Example clue words

fukuyou[2] (adverse drug effects)	*eikyou* (effect)	*okage* (virtue)
sei (reason)	*ato* (after)	*kaeru* (change)
shiyou (using)	*fukuyou* (take)	*touyo* (administration)
nomu (drink)	*tsukau* (use)	*fueru* (increase)

3.2. 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.

3.3. 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 2, rectangles indicate a phrase and arrows show a dependency parsed by CaboCha. The extraction patterns

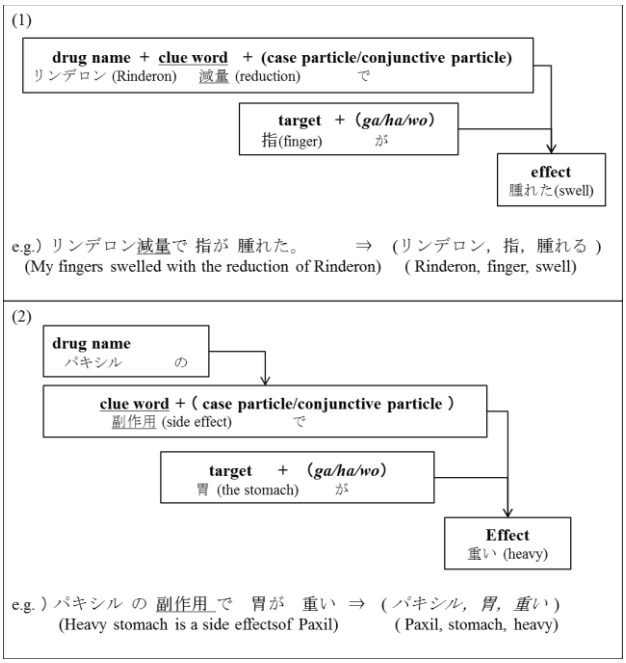

Figure 2. Pattern of extraction

were empirically developed based on syntactic structures of 181 sentences that contained drug information. For developing the patterns we also utilized rules for opinion mining from reviews.

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.

3.4. 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.

In response to the situation, our system converts as described below.

a) Combines words as the element
   Successive nouns
   A prefix and next noun
   A modifier clause which has a postpositional

[2] An italic letters shows a Japanese character.

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

particle "*no*" (“of”) (as head of sentence)

b) Make a negative expression of the element when "*nai*” (“not”) exists 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.

4. Evaluation experiment

The purpose of our experiment was to demonstrate the performance of drug information extraction using patterns with clue words 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.

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

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

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

4.1. Data

Our system extracts from newly collected 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. In order to make correct triplets of drug information, five human annotators judged whether drug information triplets were parts of newly collected 2,000 sentences containing drug names. As a result, the value of Kappa (κ) was 0.41 and if it was impossible to come up with the correct data to suit our purpose because the annotators did not consider the fact that many drugs produced no effects to be relevant drug information. In the evaluation experiment, the 340 triplets of drug information that were tagged by the first author were used as the correct data, owing to the fact that the annotators could not tag some data correctly.

4.2. Baseline systems

Shown in Figure 3 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[10], 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.

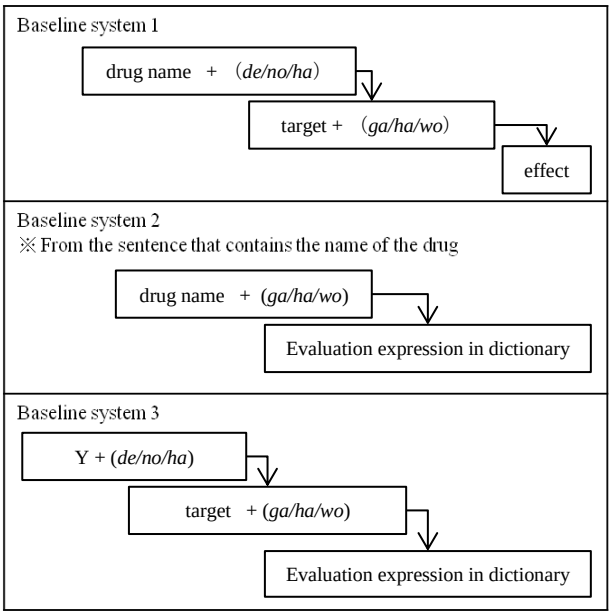

Figure 3. Baseline system extraction pattern

Table 3. 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

5. Results and discussion

The results of the evaluation are displayed in Table 3. Table 4 contains respective output examples. 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.

Table 4. Examples of the outputs

	Input	Output	Evaluate
Our system	幸いにも治療が進むにつれて自力で腎臓が動き出れて、そのまま進級ができました校長先生ありがとうでもステロイドの副作用で顔はパンパン Fortunately, with the progression of the treatment, my kidney started to move and I was able to be promoted. Thanks, principal. But my face was swollen due to a side effect of the steroids	（ステロイド，顔，パンパン） (steroids , face , swollen)	correct
	反対にリウマトレックスを飲んで翌日からしばらく吐き気と倦怠感がけっこうつよく今、痛みが出てきていると「早く月曜日になって～（月曜はリウマトレックスの服用日）」と思ったり On the contrary, fatigue and nausea are fairly strong for a while from the next day with taking Rheumatrex. Now when the pain comes out, I think "please be Monday soon" (Monday is the day for taking Rheumatrex)	（リウマトレックス，痛み，出る） (Rheumatrex , pain , comes out)	correct
	よかった頭痛については、プレドニンを投与しはじめて極度に視力が下がってきていてコンタクト It's good. Re: my headache, my sight has got drastically worse since I started taking Prednisone	（プレドニン，視力，下がる） (Prednisone , sight , worse)	correct
	いままで甲状腺ホルモンが沢山でていたが、いくら間食・暴飲暴食しても体重が増えないはずだ!!メルカゾールのおかげで病気自体はよくなってきたものの私の体がどんど Until now, I had a lot of thyroid hormone, and my weight should not increase after drinking and eating too much!! My disease is getting better thanks to Methimazole but my body rapidly (..,)	（メルカゾール，私の体，どんど） (Methimazole , my body , rapidly)	mistake
	ペガシス単独治療を終了して 1 年半以上を経過した More than one year and a half has passed since the Pegasys monotherapy finished	（ペガシス，1 年半以上，経過） (Pegasys , more than one year and a half , passed)	mistake
	イレッサ隔日服用に!今日は S 先生の外来日で肝臓関係の血液検査を診ていただきました I am now taking Iressa every other day! I had a look at the blood tests related to my liver because Dr. S took change of the outpatients	（イレッサ，!今日，診る） (Iressa , today , look)	mistake
Baseline system 1	リフレックス服用後のムズムズ感が無くなった事を伝えたあとの先生からた (…) from the doctor after I said that the itching feeling has gone after taking Reflex	（リフレックス，ムズムズ感，無くなる） (Reflex , itching feeling , has gone)	correct
	ただパキシルのほうが新薬で改良型なので吐き気の具合は改善されるかも But the condition of nausea may improve because Paxil is a new drug and an improved version	（パキシル，ほう，改良型） (Paxill , , improved)	mistake
Baseline system 2	エンブレルを開始してまだ 3 回目なのですが、痛みが激減して喜んでいま It is still only the third time since I started taking Enbrel, but I'm glad the pain has dropped sharply	（エンブレル，痛み，激減） (Enbrel , pain , drop sharply)	correct
	（左側目に傷がついていたせいで点眼薬を出してもらいました（ヒアレイン）次回の予約は 3 週間後です担当のセンセは 9 月 30 日付で異 (I got a eyedrop because my left eye was damaged (Hyalein), the next reservation is 3 weeks after, on September 30	（ヒアレイン，予約，異） (Hyalein, reservation,)	mistake
Baseline system 3	私にとってタキソテールの脱毛率は結構高い For me, the rate of hair loss with Taxotere is quite high	（タキソテール，脱毛率，高い） (Taxotere , the rate of hair loss , high)	correct
	ずーーといい天気が続いていますここは、家の前は太平洋なので冬は穏やかな日が結構多いのです関節リウマ朝食後に普段は 6 錠だけど、今日は 7 錠プレドニゾロン 2 錠、ボルタレン SR カプ It has been niiiiiiiiice weather these days. Since my house is close to the pacific ocean, there is pretty much calm weather during winter. Usually, I take 6 tablets for articular rheumatism after breakfast. However, I took 7 tablets of Prednisolone, and 2 tablets of Voltaren SR capsules today.	（ボルタレン，前，太平） (Voltaren, close , pacific)	mistake

The postpositional particles used in extraction in our system are limited. In order to investigate the effects of the limitation, we changed the conditions. When case particles and conjunctions were not conditional, the extraction number increased by 51, but the correct answers increased by only 15, and eventually, the accuracy was reduced by 11.26 points. When the system extracts the target element from the phrase depending on the phrase that contains the effect by "*ga, ha, wo*" and "*ni, mo, nimo*", the extraction number increased by 23, the correct answer increased by 5, and the precision decreased by 9.63 points. Medical information is different from information about appliances, shops, places, etc., because accuracy is more important. We found that the conditionality of our system was effective in raising accuracy.

The reasons why the recall of our system was low 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. Further consideration is required to incorporate the methods of the baseline systems in our system in order to increase recall.

6. 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.

In future works, we will change the target data to all blog articles and increase the number of clue words in order to reduce parsing errors and match the extraction patterns.

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, Sachiko Maskawa, and Mizuki Morita, "*Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter*", Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1568–1576, 2011.

[3] Kin Wah Fung, Chiang S. Jao, and Dina Demner-Fushman, "Extracting drug indication information from structured product labels using natural language processing", J Am Med Inform Assoc. 2013.

[4] 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.

[5] Corinna Cortes and Vladimir Vapnik, "*Support Vector Networks*", Machine Learning, Vol. 20, pp. 273–297, 1995.

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

[7] Emiko Shinohara, Keigo Hattori, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi, Eiji Aramaki, and Kazuhiko Ohe, "*Koubun Patahn 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.

[8] 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.

[9] 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.

[10] 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.