Chapter 7

Analyses on Text Data Related to the Safety of Drug Use Based on Text Mining Techniques

Masaomi Kimura

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/51195

1. Introduction

One of main raison d'etre of medical care should cure patients and save their lives. Drug safety has attracted attention for a long time, with an emphasis on toxicity and side effects of drugs. Additional to this, the safety of drug use is attracting increasing attention from the perspective of medical accident prevention. In order to prevent medical accidents, such as errors involving medicines, double dosage and insufficient dosage, it is necessary to ensure the proper treatment of the right medicines, namely, safety of drug use. The confirmation of usage should be one of the keys to identifying errors and prevention from misuse. Consider the case when a doctor inputs prescription data into a computerized order entry system for medicines. If the system shows him information concerning therapeutic indications, he can subsequently avoid the errors. To enable this, the order entry system requires the databases containing information on dosage regimens so that the proper usage can be verified.

The most reliable data, which can be a source of the databases, is a package insert published by pharmaceutical companies as an official document attached to its medicine. Original package inserts are, however, distributed as paper documents and unsuitable for processing by a computer system. In Japan, Pharmaceutical and Medical Devices Agency (PMDA), which is an extra-departmental body of the Japanese Ministry of Health, Labor and Welfare, has released SGML formatted package insert data. SGML is an old-established markup language, which adds metadata and structures to data by tagging, which is defined by DTD. In fact, it is difficult to leverage the data structure defined in the DTD for analysis of the data. This is because the definition of data structure is ambiguous and because the information is not well structured, namely, described by the sentences in tagged elements. This hinders the utilization of the SGML formatted package insert data, especially as a database used in computer systems that ensure the safety of medicinal usage. We should also note that the SGML version package inserts usually describe their contents as sentences, as is described in the original paper version package inserts. In order to obtain information from package insert data, we need to analyze the sentences in package insert data.

Other important sources of knowledge besides official packge inserts are practices of medical experts. One of the useful and important ways to understand what people think is to conduct

a survey in the form of a questionnaire. In particular, the freely described data included in the questionnaire responses represent an important source to let us know the real thoughts of the people. However, it is not easy to analyze such freely described data by hand, since a large number of responses are anticipated and subsequent analysis using manual counting may be influenced by the individual prejudice of the analysts involved. It is, therefore, suitable to apply a text mining approach to objectively analyze such freely described data. As readers know, text mining is an analytical technique based on data mining / statistical analysis algorithms and NLP algorithms. It has wide applicability — including clustering research papers or newspaper articles, finding trends in call center logs or blogged articles, and so on. The clustering of textual data is popular as a commonly-available method to classify data and understand their structure. Unlike such applications, however, the freely described data contained in the responses of a questionnaire have characteristics such as a small number of short sentences in each piece of data and wide-ranging content that precludes the application of clustering algorithms to classify it. In this chapter, we review the cases of application of our method to questionnaire data.

As we mentioned above, it is necessary to avoid medical accidents. In order to take a countermeasure, past cases must be investigated to identify their causes and suitable countermeasures. Medical incidents, caused by treatment with the wrong medicines, are strongly related to medical accidents occurring due to a lack of safety in drug usage. Medical incidents are the ones that may potentially become medical accidents without certain suppression factors, and tend to occur more frequently than medical accidents. Incorporating Heinrich's law, which shows the tendency of frequency and seriousness of industrial accidents, we can estimate that for every serious medical accident, there are 300 incidents and thirty minor accidents. This can be interpreted as medical accidents having many causes, most of which are eliminated by certain suppression factors, which lead to incidents, while the remaining causes lead to medical accidents. From this perspective, we can expect both medical accidents and incidents to originate from identical causes, which suggests that the analysis of data concerning incidents is valid in order to investigate the cause of medical accidents, since their occurrence frequency tends to be much larger than that of medical accidents. Though simple aggregation calculations and descriptive statistics have already been applied to drug-related medical incident data, the analyses are too simple to extract sufficient information, such as the reasons behind incidents depending on the circumstances. To ensure such analyses could be properly performed, we should apply text mining technique to the texts describing incidents.

In this chapter, we introduce the techniques that we have developed, Word-link method and Dependency-link method, and review their application to the following data:

- Package inserts
  - Application to an analysis on descriptions of dosage regimens described in package inserts of medicines
- Questionnaire data
  - Application to data obtained by nation-wide investigations based on questionnaires about the 'therapeutic classification mark' printed on transdermal cardiac patches
- Medical incident data
  - Application to incident data disclosed by Government of Japan

http://dx.doi.org/10.5772/51195

2. Method

2.1. Word-link method and dependency-link method [3]

2.1.1. Introduction

In order to determine the features of freely described data, the easiest and simplest way is to apply morphological analysis and count the number of the root (main part) of morphemes, which shows us particular words recurring frequently and suggests the nature of the themes discussed by respondents. This method, however, derives a difficult result to interpret in the case where there are several different topics contained in the entire free descriptions contained in the questionnaire responses. This is because that method can show the appearance of words but does not preserve their inter-relations. This method cannot, therefore, provide us with more in-depth information, such as how matters related to the topic are evaluated by the respondents.

Regarding the syntax tree of a sentence based on modification relationships as semi-structured data, Matsuzawa et al. [1] and Kudo et al. [2] have applied pattern mining algorithms to extract frequently appearing subtrees, namely, sub-sentences recurring frequently in plural sentences more than a specified number of times (support). These represent rigorous means to determine the pattern of sub-sentences, which preserves the co-occurrence relationships of words and their structure in sentences.

As for the freely described data written by respondents, there is no guarantee of them expressing the same opinion in sentence of the same structure. If the respondents write similar sentences but with slightly different structures, it is difficult to identify the sentences by only matching their substructures alone. In addition, we have to maintain the entire data in memory at the same time when we use the pattern mining algorithm, which prunes the substructure appearing less than the support during the process. It is preferable that the algorithm be applicable to the huge size of data to a sufficient extent to cover surveillance in the form of a large-scale questionnaire.

In this section, we, therefore, suggest a method featuring summarized description data, by initially aggregating modification relations and then limiting them to instances appearing more than the support. By connecting the resultant modification relations and finding word sequences which can be reconstituted into understandable sentences, we can expect to extract sentences which contain the main opinions of the respondents.

2.1.2. Theory

Let s_i $(i = 1 \cdots n)$ denote the sentences in freely described text data. Applying morphological analysis to s_i, we obtain a series of words $W(s_i) = \{w_1^i, w_2^i, \cdots\}$, where w_j^i denotes a word in the sentence s_i. We also define a set of dependency relations $D(s_i) = \{d_1^i, d_2^i, \cdots\}$, where d_j^i denotes a dependency relation in the sentence s_i, and their union set $D = \cup_i D(s_i)$.

For instance, if we target the two sentences, s_1="医薬品の安全性は重要だ"(the safety of drug is important) and s_2="医薬品の安全性は改善が必要だ"(the safety of drug needs improved),

- $W(s_1) = \{$ 医薬品(drug), 安全性(safety), 重要だ(important) $\}$,
- $W(s_2) = \{$ 医薬品(drug), 安全性(safety), 改善(improved), 必要だ(needs)$\}$,

- $D(s_1) = \{$ 医薬品(drug)→ 安全性(safety), 安全性(safety)→重要だ(important)$\}$,
- $D(s_2) = \{$ 医薬品(drug)→ 安全性(safety), 安全性(safety) →必要だ(needs), 改善(improved)→必要だ(needs) $\}$
- $D = \{$ 医薬品(drug)→ 安全性(safety), 安全性(safety)→重要だ(important), 安全性(safety) →必要だ(needs), 改善(improved)→必要だ(needs)$\}$.

Note that, following the linkage of $d_j^i \in D(s_i)$, we can reproduce the original sentence s_i except for the order of appearance of modifications which modify the same word. If the word w_j^i modifies another word w_k^i and the dependency relation $d^i \in D(s_i)$ is related to these words, we can define 'counterpart' functions such as

$$d^i = L(w_j^i, w_k^i) \tag{1}$$
$$w_j^i = S(d^i) \tag{2}$$
$$w_k^i = E(d^i). \tag{3}$$

The function L denotes dependency linkage between w_j^i, w_k^i and S and E returns a modifying word and a modified word respectively. For instance, as for the dependency医薬品(drug)→ 安全性(safety), $d = L($医薬品(drug), 安全性(safety)$)$, 医薬品(drug)$= S(d)$ and 安全性(safety)$=E(d)$.

Note that some relations between these functions hold as follows:

$$d^i = L(S(d^i), E(d^i)) \tag{4}$$
$$w_j^i = S(L(w_j^i, w_k^i)) \tag{5}$$
$$w_k^i = E(L(w_j^i, w_k^i)). \tag{6}$$

Let us assume the verb of the main clause is modified by other words but does not modify another word in the target language. For all dependency relations $d^i \in D(s_i)$ whose $E(d^i)$ is not the verb of the main clause of s_i, there exists another $d'^i \in D(s_i)$ which satisfies

$$E(d^i) = S(d'^i),$$

because each word but the verb of the main clause necessarily modifies other word in the sentence.

Thus, there exists $d' \in D$ satisfying $E(d) = S(d')$ for each $d \in D$, if d is not the verb of the main clause of the original sentences $\{s_i\}$. Since $D(s_i) \subset D$ because of the definition of D, we can find a series of modification relations which satisfy the Eq.2.1.2 in D and reproduce all the original sentences $\{s_i\}$ by following their linkage.

However, rather than all sentences, we are only interested in the sentences described by plural respondents. If the same sentences appear η times, the dependency relations in the sentences will also recur (more than) η times. Let us define a 'support' function:

$$supp_D(d) = card\{s_i \mid d \in D(s_i)\},$$

http://dx.doi.org/10.5772/51195

where 'card' denotes the cardinality of a set. The above statement can be described via $supp_D(d)$ as follows: if there are η sentences, which have the same dependency structure as s_i , the number of sentences is equivalent to η, which contains $d_k^i \in D(s_i)$. Thus the following inequality holds for each $d_k^i \in D(s_i)$

$$supp_D(d_k^i) \geq \eta.$$

Therefore, If we limit D to the set with the constraint of Eq.2.1.2:

$$D^\eta = \{d \mid d \in D, supp_D(d) \geq \eta\},$$

each modification relation in sentences with the same dependency structure, namely more than η times, is a member of D^η. These dependency relations satisfy the same relation as Eq.2.1.2, though, in general, we cannot necessarily expect the existence of the dependency relation $d \in D^\eta$ such that $E(d) = S(d')$ for each $d' \in D^\eta$. We can therefore expect to find sentences described by plural respondents and with an equivalent dependency structure by following the linkage of dependency relations in D^η, which satisfies the relation Eq.2.1.2. (We call this method using a series of dependency relations the 'word-link method'.)

In fact, we should be aware that the extraction of a series of dependency relations in D^η satisfying Eq.2.1.2 is a necessary condition to find such sentences and the co-occurrence of dependency relations is not preserved in this operation. In other words, the elements of D^η, d and d', which satisfy the relation $E(d) = S(d')$, do not necessarily appear in the same sentence. In order to ensure the co-occurrence of dependency relations, it is necessary to confirm that the dependency relations d and d' satisfying $E(d) = S(d')$ are included in the same sentence. If more sentences exist than the support, which contains a series of dependency relations satisfying $E(d) = S(d')$, we can conclude that the sentences are written by more respondents than the number preliminarily determined. Taking the calculation cost and the degree of freedom of expression into account, we relax the above restriction as follows:

1. Firstly, find the pairs of dependency relations $d, d' \in D$ satisfying $E(d) = S(d')$, both of which are contained in the same sentence. Let $d \to d'$ denotes such a pair of dependency relations (First step).
2. Next, find the two pairs $d \to d'$ and $d' \to d''$, where the dependency relationship d' in both pairs is identical. If such pairs exist, we presume there is a link connecting these pairs (Second step).
3. Finally, follow the linkages of such pairs which appear more than η' times and reproduce sentences (Third step). η' is the threshold to limit the lower boundary of the number of appearances.

In this method, each of two pairs of dependency relations $d \to d'$ and $d' \to d''$ contains a common pair of words $E(d) = S(d')$ and $E(d') = S(d'')$, which appears in the same sentence. Since the variations of the structures of descriptions related to common opinions in a set of questionnaire data tend to be small, such overlap of words is (at least empirically) sufficient

to approximately reproduce sentences summarizing original sentences. (We call this method using the series of the pairs of modification relations the 'dependency-link method'.)

In addition, our method helps us find the sentences which have similar dependency structures. We usally visualize the result as a graph structure, whose nodes denote modifying or modified words and edges denote dependency relationships between the words. We can expect that such sentences are placed in the same graph structure since they share the same words and the similar dependency relations.

3. Application

3.1. Analysis on descriptions of dosage regimens in package inserts of medicines [4]

To prevent medical accidents, such as mix-ups involving medicines, double dosage and insufficient dosage, it is necessary to ensure the proper treatment of the right medicines, namely, 'safety of usage' of medicines.

There occurred, in some Japanese hospitals, fatal accidents due to mix-ups involving a steroid, Saxizon, with a similarly-titled medicine, Succine, which is a muscle relaxant. There are two conceivable ways to avoid such accidents, one of which is to prevent the naming and use of medicines resembling other medicines in their name, both in terms of appearance and sound. Another method is to confirm the medicine by checking the actual usage based on their dosage regimens. Though the former method can be realized by utilizing a name checking system provided by the Japan Pharmaceutical Information Center or making a rule to adopt medicines which have confusing names, the accident is known to have occurred despite the existence of a rule to reject Succine due to its confusing name.

This suggests to us that the latter, namely the confirmation of usage, should be the key to identifying error. Consider the case when a doctor inputs prescription data into a computerized order entry system for medicines. If the system shows him information concerning therapeutic indications, he can subsequently avoid mix-ups of medicines such as the case in question. To enable this, the order entry system requires a database containing information on dosage regimens so that the proper usage can be verified.

As is described in Introduction in this chapter, the structure of the portion of dosage regimens in package insert data does not achieve sufficiently fine granularity to enable its effective utilization in a computer system, such as the order entry system mentioned above. In this section, we show the method to find the description patterns of the sentences in the dosage regimen portion of the SGML formatted package inserts data. Based on this result, we also propose the data structure of dosage regimen information, which will be the basis of a drug information database to ensure safe usage.

The target data in this section is the SGML formatted package insert data of medicines for medical care, which can be downloaded from the PMDA web site. Since we need the list of medicines to retrieve the data, we utilize the standard medicine master data (the version released on September 30, 2007), which is provided with The Medical Information System Development Center (MEDIS-DC). Using the master data, we obtained 11,685 SGML files, which are our target data.

http://dx.doi.org/10.5772/51195

The part of dosage regimens contain 'detail' elements. They describe information concerning dosage regimens as sentences and are suitable to apply a text mining technique in order to find potential meta data of dosage regimens.

We applied the word-link method to descriptions in 'detail' elements concerning the dosage regimens in each SGML package insert. Since, as a minimum, dosage, administration and adaptation diseases will differ for each medicine, with a considerable scope of expression, our original method, whereby attempts are made to find patterns, including the use of nouns, might result in a failure to find the common sentences. We thus extend it to determine the tendency for the co-occurrence of nouns and particles (parts of speech which play roles similar to prepositions in English) and extract structural patterns except for noun variations. The analytical steps are as follows:

1. We retrieve sentences in the 'detail' elements and apply dependency analysis to them.
2. If the segment in the dependency contains a noun, we differentiate the latter from the segment. The resultant characters are expected to be particles, hence we name a 'particle candidate' in this paper.
3. We aggregate nouns that appear in segments including each particle candidate and find the characteristics of the particle candidates in use. We call the part of the segment obtained by removing a particle segment the 'main part of segment'.
4. We replace the found nouns with a symbol such as '○○○' in order to mask them, and apply the word-link method. If there are certain rules governing the way in which particles should be used, this method extracts the common structures of sentences and suggests us the idea of data items, for which descriptions must be converted into a structured data form.

value	ratio	%	freq.
		53.67	155411
を	(to)	8.67	25094
に	(to)	6.74	19516
として	(as)	4.67	13512
の	(of)	3.97	11506
により	(depending on)	3.31	9578
には	(for)	3.0	8696
は	(is)	2.12	6143
$null$		2.08	6011
,		1.84	5331
が	(is)	1.54	4448
と	(with/and)	1.15	3333
で	(with/by)	0.76	2195
まで	(to/untill)	0.4	1163
な	(like)	0.36	1053
も	(too/also)	0.28	817
後	(after)	0.28	813
では	(as for)	0.24	688
から	(from)	0.22	634
までと	(until)	0.2	593

Figure 1. The particle candidates of segments included in the 'detail' elements.

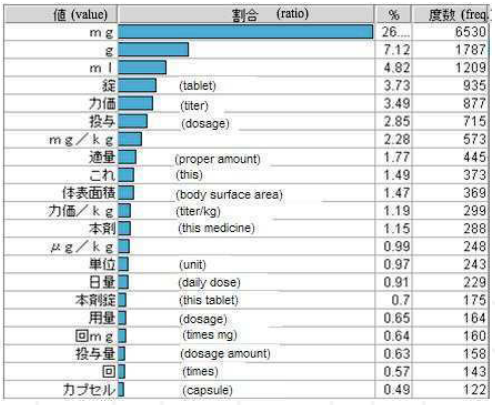
Fig. 1 shows the distribution chart of particle candidates with their frequencies. First, we investigate the nature of the nouns involved in the segments containing the particle candidates appearing frequently in the sentences of dosage regimens. Fig. 1 indicates that the particle candidate of more than 50% of the segments is a null character, namely the segments contain only their main part. Since the targets in Fig. 1 are all segments contained in sentences of dosage regimens, they involve not only nouns but also other part of speech

such as verbs. The particle candidate of segments whose main word is not a noun is expected to be a null character. In the following analysis, we thus exclude segments whose main word is not a noun.

Fig. 2 shows nouns in the segments whose particle candidate is a null character. This indicates that such segments contain information about units of administration, '日' (days), '回' (times), 'mg', the manner of administration, '適宜' (arbitrarily), '通常' (usually), and the condition of age such as '年齢' (age) and '成人' (adult) and so on.

value	ratio	%	freq.
日	(day)	17.84	15125
適宜	(proper extent)	12.52	10620
回	(times)	11.37	9640
通常	(ordinary)	10.03	8503
年齢	(age)	7.31	6200
ｍｇ		4.89	4148
通常成人	(ordinary adult)	3.34	2834
成人	(adult)	2.29	1942
体重	(weight)	1.65	1403
こと	(thing)	1.36	1154
ｇ		1.17	994
日量	(daily amount)	1.07	904
ｍｌ		0.85	724
本剤	(this medicine)	0.78	659
錠	(tablet)	0.76	648
症状	(symptom)	0.56	479
場合	(case)	0.56	473
時間	(time)	0.54	458
日間	(during ... days)	0.42	356
力価	(titer)	0.41	348

Figure 2. The nouns whose segment has a null character as the particle candidate.



値 (value)	割合 (ratio)	%	度数 (freq.
ｍｇ		26....	6530
ｇ		7.12	1787
ｍｌ		4.82	1209
錠	(tablet)	3.73	935
力価	(titer)	3.49	877
投与	(dosage)	2.85	715
ｍｇ／ｋｇ		2.28	573
適量	(proper amount)	1.77	445
これ	(this)	1.49	373
体表面積	(body surface area)	1.47	369
力価／ｋｇ	(titer/kg)	1.19	299
本剤	(this medicine)	1.15	288
μｇ／ｋｇ		0.99	248
単位	(unit)	0.97	243
日量	(daily dose)	0.91	229
本剤錠	(this tablet)	0.7	175
用量	(dosage)	0.65	164
回ｍｇ	(times mg)	0.64	160
投与量	(dosage amount)	0.63	158
回	(times)	0.57	143
カプセル	(capsule)	0.49	122

Figure 3. The nouns whose segment has a particle candidate 'を'. (top 20)

We outline the nouns in the segments, including each particle segment, as follows:

- Fig. 3 shows nouns in the segments, including 'を' as a particle segment. We can see that they express amounts of medication such as 'mg', '錠' (tablets) and '力価' (titers).
- The nouns in the segments whose particle segment is 'に' (at/to) are shown in Fig. 4, which shows that the particle segments tend to be used with frequency-related words such as '回' (times) and '数回' (sometimes), and concerning the timing of administration, such as '食間' (inter cibos) and '就寝前' (before bedtime), administration site such as '静脈内' (in a vein).

http://dx.doi.org/10.5772/51195

値 (value)	割合 (ratio)	%	度数 (freq.)
回	(times)	23.2	4527
症状	(symptom)	5.46	1066
食間	(inter cibos)	4.7	918
就寝前	(before bedtime)	3.28	641
静脈内	(in the veins)	3.21	626
必要	(necessity)	2.53	494
患部	(affected area)	2.18	425
緩徐	(moderation)	2.08	406
ｍｌ		1.92	375
食後	(after meal)	1.9	370
回食後	(times after meal)	1.43	280
成人	(adult)	1.13	221
数回	(some times)	1.01	197
回患部	(times affected area)	0.92	180
時間ごと	(each ... hour)	0.87	169
経口投与	(oral administration)	0.81	158
筋肉内	(intramuscular)	0.81	158
週間ごと	(each ... weeks)	0.75	147
年齢症状	(age and symptom)	0.67	131
食前	(before meal)	0.61	119

Figure 4. The nouns whose segment has a particle candidate 'に' (at/to).(top 20)

値 (value)	割合 (ratio)	%	度数 (freq.)
原則	(principle)	2.9	392
クラリスロマイシン	(clarithromycin)	1.18	159
日量	(daily dose)	1.17	158
ファモチジン	(famotidine)	1.04	141
回アシクロビル	(times acyclovir)	1.03	139
回アルファカルシドール	(times alfacalcidol)	0.97	131
フルオロウラシル	(fluorouracil)	0.94	127
維持量	(maintainance dosage)	0.94	127
シメチジン	(cimetidine)	0.85	115
ドンペリドン	(domperidone)	0.78	106
ドライシロップ	(dry syrup)	0.74	100
ケトチフェン	(ketotifen)	0.72	97
アモキシシリン	(amoxicillin)	0.68	92
レボホリナート	(Levofolinate)	0.65	88
ランソプラゾール	(lansoprazole)	0.64	86
オメプラゾール	(Omeprazole)	0.62	84
ジクロフェナクナトリウム	(diclofenac sodium)	0.62	84
メシル酸ナファモスタット	(nafamostat mesilate)	0.61	83
プランルカスト水和物	(pranlukast hydrate)	0.61	82
ラニチジン	(ranitidine)	0.61	82

Figure 5. The nouns whose segment has a particle candidate 'として' (as). (top 20)

値 (value)	割合 (ratio)	%	度数 (freq.)
(adult) 成人		49.28	3976
(case) 場合		13.94	1125
(ordinary adult) 通常成人		8.96	723
(child) 小児		7.11	574
(severe infection) 重症感染症		1.72	139
(elder people) 高齢者		0.71	57
(improvement) 改善		0.66	53
(dissolution and dilution) 溶解希釈		0.57	46
(patient) 患者		0.56	45
(case) 症例		0.48	39
(ordinary child) 通常小児		0.43	35
(hepatic disease) 肝疾患		0.37	30
(less than ... years old) 歳未満		0.36	29
(after) 後		0.33	27
(objective) 目的		0.33	27
(severe hypertension) 重症高血圧症		0.33	27
(astriction) 便秘		0.32	26
(insomnia) 不眠症		0.29	23
重症難治性感染症	(severe refractory infection)	0.29	23
(baby and child) 幼小児		0.27	22

Figure 6. The nouns whose segment has a particle candidate 'には' (for). (top 20)

値 (value)	割合 (ratio)	%	度数 (freq.)
症状	(symptom)	80.27	7688
年齢症状	(age and symptom)	7.49	717
体重	(weight)	2.14	205
目的	(objective)	1.41	135
年齢体重症状	(age, weight and symptom)	0.67	64
体質	(diathesis)	0.56	54
病型	(pattern of disease)	0.55	53
経口投与年齢症状	(oral administration, age and symptom)	0.41	39
年令症状	(age and symptom)	0.41	39
疾患	(disease)	0.39	37
こと	(thing)	0.32	31
程度	(extent)	0.31	30
年齢症状等	(age and symptom etc.)	0.28	27
尿量	(urinary volume)	0.28	27
患者	(patient)	0.25	24
抗悪性腫瘍剤	(anticancer drug)	0.24	23
状態	(state)	0.22	21
次式	(next equation)	0.2	19
生理食塩液	(normal saline)	0.19	18
大きさ	(size)	0.18	17

Figure 7. The nouns whose segment has a particle candidate 'により' (depending on). (top 20)

- The particle segment 'として' (as) is included in the segments whose main words are nouns, as shown in Fig. 5. Besides the nouns for the formulaic phrases, '原則として' (as a rule), '(1) 日量として' (as a daily dosage) and '維持量として' (as a maintenance dosage), the other nouns shown in the figure represent active ingredients of medicines.
- Fig. 6 shows nouns in the segments including the particle segment 'には' (for). This mainly contains nouns showing an object person such as '成人' (adult), '小児' (child) and '高齢者' (elder person). It also shows the name of symptoms such as '重症感染症' (severe infection) and '肝疾患' (hepatic disease).
- In Fig. 7, segments whose particle candidate is 'により' (depending on) tend to contain the word '症状' (symptom). In this figure, we can also read words such as '体重' (body weight), '年齢' (age), '目的' (objective) and so on. This results and the meaning of the particle candidate suggest that these segments show the condition to adjust a dose.

値 (value)	割合 (ratio)	%	度数 (freq.)
増減する	(increase and decrease)	17.23	10115
経口投与する		9.72	5708
(administer orally) する	(do)	5.81	3409
分ける	(split)	3.64	2135
応じる	(respond)	3.61	2122
投与する	(administer)	2.85	1676
用いる	(use)	2.57	1512
分割する	(devide)	2.39	1402
分割経口投与する	(take orally in fractional	2.26	1330
(dissolve) 溶解する	amounts)	1.75	1026
増量する	(increase)	1.73	1018
できる	(be possible)	1.67	983
点滴静注する	(infuse intravenously)	1.65	970
使用する	(make use of)	1.44	843
注射する	(inject)	1.41	827
かける	(source)	1.4	822
3回経口投与する	(administer orally thrice)	1.29	757
開始する	(begin)	1.1	645
塗布する	(apply)	1.04	608
行う	(perform)	1.01	596

Figure 8. The verbs included in 'detail' elements describing dosage regimens.(top 20)

Based on the results shown above, we can find the tendency of contents in the segments including each particle segment. We replaced each segment containing nouns with the symbol '○○○', and applied the word-link method to the replaced sentences. Fig. 8

http://dx.doi.org/10.5772/51195

shows the verbs used in the sentences of dosage regimens. To absorb the difference in verb expressions, we replace verbs of similar meanings with a representative verb. For instance, the verbs, '経口投与する' (dose orally) and '点滴静注する' (drip-feed intravenously) have analogous meanings in terms of medication and are hence consolidated into a single verb. In this paper, to enhance comprehension, we consolidated them into '投与・使用する' (administrate/use). Moreover, we consolidated the verbs that mean increase or decrease into '増減する' (escalate) and replaced the verb '分割する' (divide) with '分ける' (split).

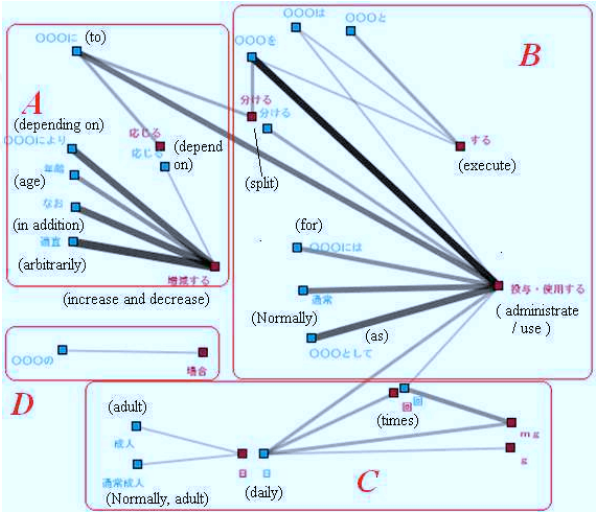


Figure 9. The result of the word-link method applied to 'detail' elements (the links show co-occurrence more than 1149 times). Blue nodes denote modifying words and red nodes denote modified words.

Following this consolidation, we applied the word-link method and obtained sentence structures based on dependency relationships. Fig. 9 shows the links of dependency relationships appearing more than 1149 times. Based on this figure, we can read the following contents:

- Increase or decrease according to conditions such as indication (disease) and age (Part A in Fig. 9).
- Dosage based on the information concerning the administration site, frequency, object person, symptoms, amount of medication and (the amount of) active gradients (Part B).
- Daily dosage (Part C) and description of conditions (Part D)

Based on these and the fact that verbs indicate the method of administration, we can see that the data structure to describe dosage regimens needs the following items:

- Indication (disease)
- Objective person
- Administration site
- Amount of medication

No.	Questions (Originally Japanese)	Respondents	Num of responses
Q1	Why did you select the transdermal patch?	Doctors, Pharmacists, Nurses	737
Q2	What are the preventive measures to avoid medical accidents related to the transdermal patch?	Doctors, Pharmacists, Nurses	2115
Q3	What is your opinion of the cardiac transdermal patch?	Patients	529
Q4	Have you ever been asked by patients about the transdermal patch with the therapeutic classification mark on it?	Pharmacists, Nurses	533

Table 1. The free description part of the questionnaire concrening the therapeutic classification mark printed on a cardiac transdermal patch.

- Amount of active gradient
- The way of administration
- Frequency
- Conditions of increase or decrease

3.2. A questionnaire concerning the therapeutic classification mark printed on a cardiac transdermal patch [5]

In certain hospitals in Japan, medical accidents have occurred, whereby patients suffering from lung ailments and those suffering from heart disease were mixed up and operations were performed without any modification. It is known that the incident happened because a cardiac transdermal patch was placed on the body of the heart disease sufferer, which indicated when the patients were delivered. If surgeons had known what the patch signified, they would have avoided making a mistake with the surgery. To prevent recurrences, the pharmaceutical company marketing the patches voluntarily printed a 'therapeutic classification mark' on them. The 'therapeutic classification mark' is a security feature linked to the use of the drug and shows that the patch is a cardiac medicine. We applied our method to the free description part of a questionnaire, which is conducted as a nationwide investigation into the 'therapeutic classification mark' printed on isosorbide dinitrate transdermal patches. The respondents were doctors, pharmacists, nurses and patients and the number of respondents and the questions asked are listed in Table 1.

Table 2 lists the resulting sentences for the dependency-linking method($\eta' = 3$), where we filled postpositions and implemented classification by respondent and topic. We only presented representative sentences in the content columns where there are many sentences with similar meanings.

Respondents	[A typical sentence (translated)] Examples of sentences originally obtained by the method.
Doctors, Pharmacists, Nurses	**[It is usable for the patients who have a difficulty in taking the medicine orally.]** 飲めない患者に使用可 経口が難しい患者に使用可
Doctors, Nurses	**[Easy to use.]** 簡単
Pharmacists	**[There are few burdens for patients.]** 患者の負担軽減
Nurses	**[I do not know well.]** よくわからない
Doctors	**[The number of oral drugs decreases.]** 経口薬の数が減る 経口薬の種類が減る 経口の量が減る **[The medicine works slowly.]** 効果が穏やか / 効果が弱い **[For the hope/ease of patients]** 患者の希望 / 患者の安心 **[Mental effects]** 心理的効果

Table 2. The resultant sentences obtained by the dependency-link method for Q1. In this table we show the typical sentences translated in English with some examples of original sentences in Japanese.

The table shows that all medical experts prioritized reducing the load of patients as the reason for selecting the transdermal patch, since it could be used by patients who were unable to take medicines orally. In addition, this shows that doctors and nurses focused on the ease of use and that doctors also prioritized the effect of the medicine.

The following is a summary of the results for Q2 -Q4 obtained by the dependency-link method:

For Q 2, the result shows that medical experts appreciated the name of the medicine and the therapeutic classification mark printed on the patch in order to prevent medical accidents and considered it necessary to have a space for the date. The doctors also required a patch that was much smaller and that changed color depending on the amount of time having elapsed. The nurses focused on the behavior of patients, while the pharmacists emphasized the widespread need for awareness regarding correct use of the medicine.

The result of Q 3 shows numerous patients' opinions concerning the medicine, skin symptoms, mentality, and the site of the patch. We can also see that patients in their 40s and 50s mainly commented on skin symptoms, although those in their 60s to 80s covered all these opinions. This suggests that the younger generation focused on the functions of the medicine, while older patients focused on other factors, like ease of mind.

For Q 4, we obtained a result showing that patients asked nurses and pharmacists questions about where to place the patch and how to use it. Nurses also asked questions concerning the effect of the medicine, while pharmacists asked about displays on the patch or packaging and when to use it. This suggests that patients expect nurses to tell them about the efficacy of the medicine and pharmacists to tell them about usage.

The result clarifies that opinions differed depending on the viewpoints of the respondents, although they all wanted to use the same medicine safely. This meant that it is necessary to collect and analyze people's opinions from various backgrounds to ensure drugs are being used safely.

3.3. Incident data related to the safety of drug use [6]

The target data were reports of medical near-miss cases related to medicines and collected by the surveys of the Japan Council for Quality Health Care, which is an extra-departmental body of the Japanese Ministry of Health, Labor and Welfare. We analyzed 858 records from the 12th - 14th surveys, whose data attributes are shown in Table 3. This is because they contain free-description data such as 'Background / cause of the incident' and 'Candidates of counter measures'. Applying text mining to such data required the deletion of characters such as symbols and unnecessary line feed characters. We must also standardize synonyms, since it is difficult to control by making respondents use standard terms to reduce the number of diverse expressions. For this reason, we standardized the words using the dictionary prepared for this analysis.

Day of the week
Weekday or holiday
Time
Place
Department
Content of incident
Psychosomatic state of the patient
Job title
Experience (year/month)
Affiliation (year/month)
Medical benefit class
Nonproprietary name
Name of wrong drug
Dosage form of wrong drug
Effect of wrong drug
Name of right drug
Dosage form of right drug
Medical benefit of right drug
Discussed cause
Concrete descriptions of the incident
Background/cause of the incidents
Candidates of counter measures
Comment

Table 3. Data attributes of records corrected by 12th - 14th surveys.

http://dx.doi.org/10.5772/51195

3.3.1. background/cause of incidents

We applied the Word-link method to data in the field 'background/cause of incidents' in order to determine the concrete information concerning the cause of incidents. The method was applied by occupation to determine the difference in backgrounds and the causes of incidents depending on the job title. We fixed the value of each η so as to make a resultant graph understandable for us. Figure 10 and Fig. 11 show the result of nurses' and pharmacists' comments, respectively. Both figures contain the common opinions, namely, 'the problem of the checking system of the protocol and the rule' (A) and 'confirmation is insufficient' (B), nurses point out 'the systematic problem of communication'(C) and pharmacists 'the problem of adoption of medicines' (C'). We can see that, though B arises due to individual faults, A, C and C' are systematic problems.

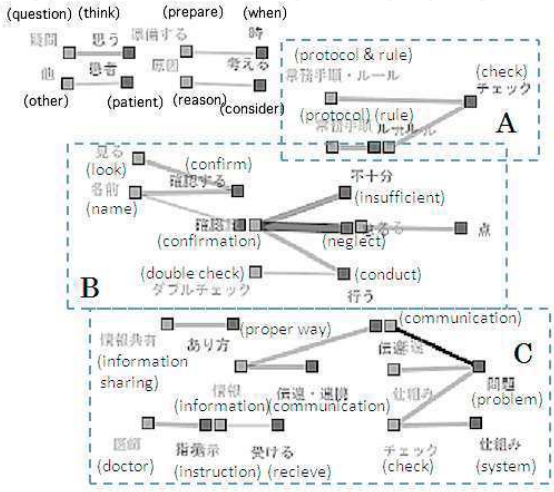


Figure 10. The backgrounds and causes of incidents caused by nurses ($\eta = 4$).

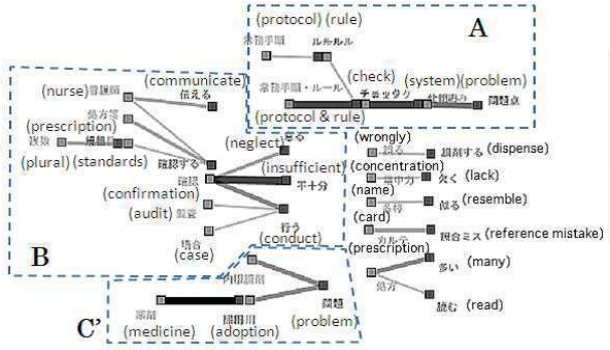


Figure 11. The backgrounds and causes of incidents caused by pharmacists ($\eta = 3$).

3.3.2. Countermeasures

We applied Word-link method to the field 'Candidates of countermeasures' to summarize the nurses' and the pharmacists' opinions concerning the countermeasures to prevent the incidents. Fig. 12 is the summary of the counter measures described by nurses, and suggests that there are many opinions stating '(it is necessary to) instruct to confirm and check', 'make a speech' and 'ensure confirmation'. Fig. 13 shows the summary of the countermeasures proposed by pharmacists. This explains that, besides the confirmation and audit, it is also necessary to attract (pharmacists') attention and to devise ways of displaying medicines such as labels.

Compared with the both results, except for the pharmacists' opinion concerning the innovation of labels, only few opinions exist on the countermeasures related to the system of the medical scenarios. This suggests that the medical experts such as nurses and pharmacists tend to try to find solutions to problems within themselves. To solve the structural problems of medical situations, it is important not only to promote the efforts of each medical expert, but also to strive to improve the organization to which they belong. It is also desirable for them to be aware of the importance of organizational innovation, and to combat the systematic error.

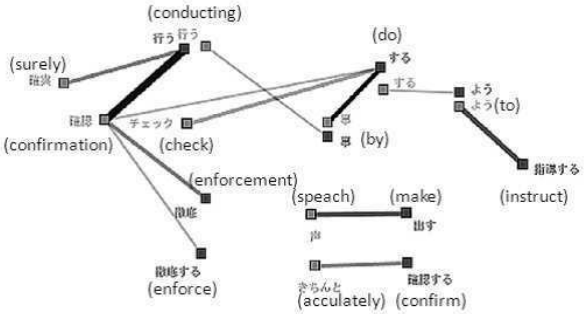


Figure 12. The countermeasures of incidents caused by nurses. ($\eta = 5$)

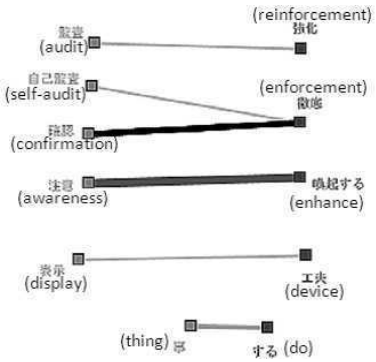


Figure 13. The countermeasures of incidents caused by pharmacists. ($\eta = 4$)

http://dx.doi.org/10.5772/51195

4. Discussion

4.1. Methods

The three analyses suggest that our method can be a powerful tool to extract the parts of sentences that commonly appear in original sentences. The target data have been Japanese sentences. Let us discuss whether our method is applicable to the data in the other language, English. As we introduced in Section 2.1, Word-link method and Dependency-link method utilize dependency relationships in target sentences. One of the representative dependency parsers for English sentences is Stanford parser [7–9], which provides us with the dependency relationships in Stanford Dependencies format. In principle, it enables us to perform our method.

The difference between Japanese and English data comes from the followings:

- Directions of dependency relationships. The dependency relationships in a Japanese sentence always have forward direction, whereas the relationships in an English sentence can have both forward and backward direction. Let us show an example that illustrates this. The Japanese sentence 'ジョンが太郎に話した' corresponds to the English sentence 'John talked to Taro'. In the both sentences, there exist dependency relationships, "ジョン(John) → 話す(talk)" and "太郎(Taro) → 話す(talk)". We should note that both 'ジョン'(John) and '太郎'(Taro) also appear prior to the verb '話す'(talk) in the Japanese sentence. This coincidence of order helps us to suggest the sentences that frequently appear in original data. [1] However, in the English sentence, the noun 'Taro' follows the verb 'talked'. Though this helps to distinguish a subject and an object, it does not preserve the order of words that appear in original sentences. Because of this, as for the dependency relationship between an object and a verb, we should swap their order (e.g. 話す(talk) → 太郎(Taro)) to reproduce summarizing sentences.
- Treatment of a relative pronoun. In English sentences, we frequently use a relative pronoun. It essentially requires reference resolution to identify an antecedent that is modified by the relative pronoun. Reference resolution often requires semantics of words and the knowledge related to them. Because of this, it is currently a difficult problem to find a right antecedent. In contrast, Japanese language does not have a relative pronoun. The relationship between a relative clause and its antecedent is built in normal modification relationships. Therefore, Japanese sentences do not cause the difficulty that originates from a relative pronoun.
- Zero pronoun. In Japanese language, we often omits a subject in a sentence. Such omission is usually called as 'zero pronoun'. In contrast, a subject in an English sentence is seldom omitted. This fact tells us that we can expect the patterns that include subjects in English sentences. If there are only the patterns without subjects, this indicates no definite subjects that appear in the target sentences. However, as for Japanese sentences, we cannot necessarily obtain information about subjects and may have to guess them based on the semantics of words included in the obtained patterns.

[1] Of course, if you need to distinguish which is a subject or an object, you should focus on particles as we did in Section 3.1.

4.2. Application

In this subsection, let us briefly review related works and discuss text mining applied to the description data related to medical safety.

4.2.1. Package inserts

U.S. Food and Drug Administration [14] also defines a specification of a package insert document markup standard, Structured Product Labeling (SPL), and . This is similar to SGML formatted package inserts disclosed by PMDA. Thus, in this chapter, we identify SPL with package inserts.

Recently, there emerge several studies which analyze descriptions in drug package inserts. Let us review some of them.

Duke et al. [10, 11] developed a tool, SPLICER, which utilized natural language processing to extract information from SPLs. It parses SPL by identification of target parts, removal of XML tags and extraction of terms. It also identify sysnonymns of the extracted terms by mapping them to medical dictionary, MedDRA. In their study, they applied their tool to quantitatively show the "overwarning" of adverse events in the package inserts of newer and more commonly prescribed drugs. They also showed that recent FDA guide lines do not succeed in reducing overwarning.

Bisgin et al. [12] applied a text mining method, topic modeling, to package insert data. A topic modeling method, latent Dirichlet allocation (LDA), explores the probabilistic patterns of 'topics', implicitly expressed by words in documents. They identified topics corresponding to adverse events or therapeutic application. This enabled them to identify potential adverse events that might arise from specific drugs.

Richard et al. [13] applied machine learning techniques to package insert data. It is a trial to automatically identify pharmacokinetic drug-drug interaction based on unstructured data. They created a corpus of package inserts, which is manually annotated by a pharmacist and a drug information expert. Using the corpus data as a training set, they evaluated the accuracy of identification and obtained F-measure of 0.8-0.9.

The number of the studies that deal with adverse events seems to be much more than the ones that deal with safety of drug usage. For the purpose of finding adverse events, package inserts are just one of text sources. Other sources are academic papers or Medline abstracts. We expect that there emerge more studies from the various viewpoint of safety to utilize package insert data.

4.2.2. Questionnaire data

There are many studies where text mining approach is applied to questionnaire data. However, as for application in the area of medication, there are only a few studies. This might be because analysts tend to take a traditional approach, manual reading, because it captures the written information more precisely than text mining. However, it is obviously time and cost consuming.

Suzuki et al. [15] applied a text mining technique to questionnaire data about clinical practice pre-education conducted to pharmacists, providers of clinical practices. Their method

was correspondence analysis between keywords appearing in sentences and attributes of respondents, such as a type of their affiliation and their profession. As a result, they obtained the tendency that mentors in hospitals feel anxious about mismatch between learning contents and real situation.

4.2.3. Medical incident data

Malpractice reduction is one of important themes of medical safety. A lot of governments or institutions construct incident reporting system and analyze the collected report data to find knowledge therein.

Kawanaka et al. [16, 17] utilized Self Organizing Map (SOM) to make a map expressing the relationships of sentences in incident report data. They calculated the co-occurrence possibility of keywords in sentences and defined a characteristic vector for each keyword. They also defined a vector to characterize a report by summing up the vectors whose corresponding keywords appear in it. They input a vector for each report to SOM algorithm. As a result, they found two clusters of reports, the former of which is summarized as "Forget of inscription to medication note" and the latter is as "Forget of administration of medicine taken before sleep". Based on this technique, they also proposed an incident report analysis system.

Baba et al. [18] proposed a method to analyze the co-occurrence relation of the words that appear in the medical incident reports using concept lattice.

Classification is a start point to analyze incident reports. Empirically, the incident types seem to obey Zipf's law. This makes it difficult to classify reports by naive application of clustering algorithms, because they generate too many small-size clusters or a large-size cluster of . If we target major incidents, the better strategy to understand reports is to focus on relative large-size clusters and to summarize the reports in them. However, one should also note that there exist important but less frequently occurring cases. Thus, it is expected to introduce a parameter to measure importance and use it to narrow down clusters to focus on.

All of the above studies suggest that text mining studies tend to focus on words not syntactic structures. Remember that stochastical approach and data mining assume table-type structured data.This might be the reason why it is more difficult to analyze syntactic structures than words. However, as Richard et al. pointed out the importance of the use of syntactic information [13], syntactic structures include information much richer than just a collection of words. They also provide us with easier interpretation of results. This is a basis of the strategy of our method.

5. Conclusion

In this chapter, we introduced the text mining method to analyze text data such as documents and questionnaire response data, and reviewed the studies where we used the method.

Our method utilizes syntactical information of target sentences. We extract a dependency relations from each sentence and restrict them to the ones that appear more than frequency threshold. Connecting common words in the resultant dependencies produces the patterns that contain the frequently appearing portions of sentences. We reviewed the study where we applied the method to drug package inserts, questionnaire data and medical incident

reports. We discussed the consideration points to apply our method to English sentences. We also introduced the related works and discussed their tendency.

Though an analysis on medical safety data is important, most of the data are untouched to be analyzed. It is expected that not only text mining techniques are developed but also they are applied to medical safety data.

Author details

Masaomi Kimura

Shibaura Institute of Technology, Japan

6. References

[1] Matsuzawa, H. (2001) Mining Structured Association Patterns from Large Databases, *Transactions of Information Processing Society of Japan*, Vol.42, No.SIG 8(TOD 10), pp.21-35.

[2] Kudo, T., Yamamoto, K., Tsuboi, Y., Matsumoto, Y. (2002) Mining Syntactic Structures from Text Database, *IPSJ SIG Notes. ICS*, Vol.2002, No.45(20020523), pp.139-144.

[3] Kimura, M. (2009) The Method to Analyze Freely Described Data from Questionnaires *Journal of Advanced Computational Intelligence and Intelligent Informatics* , Vol.13 No.3 pp.268-274.

[4] Kimura, M.; Okada, K.; Nabeta, K.; Ohkura,M.; Tsuchiya, F. (2009) Analysis on Descriptions of Dosage Regimens in Package Inserts of Medicines, In: *Human Interface and the Management of Information. Information and Interaction*, Vol.5618 pp.539-548.

[5] Kimura, M.; Furukawa, H.; Tsukamoto, H.; Tasaki, H.; Kuga, M.; Ohkura,M.; Tsuchiya, F. (2005) Analysis of Questionnaires Regarding Safety of Drug Use, Application of Text Mining to Free Description Questionnaires, *The Japanese Journal of Ergonomics*, Vol.41 No.5 pp.297-305.

[6] Kimura,M.; Tatsuno,K.; Hayasaka,T.; Takahashi,Y.; Aoto,T.; Ohkura,M.; Tsuchiya,F.(2007) The Analysis of Near-Miss Cases Using Data-Mining Approach. *Human-Computer Interaction. HCI Applications and Services* pp.474-483, Beijing.

[7] Klein, D. & Manning, C. (2003a) Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

[8] Klein, D. & Manning, C. (2003b). Fast Exact Inference with a Factored Model for Natural Language Parsing. In: *Advances in Neural Information Processing Systems 15*, Cambridge, MA: MIT Press, pp. 3-10.

[9] Marneffe, M.C.; Bill MacCartney, B.; Manning, C.(2006) Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC*.

[10] Duke, J. & Friedlin, J. (2010) ADESSA: A Real-Time Decision Support Service for Delivery of Semantically Coded Adverse Drug Event Data. *AMIA Annu Symp Proc. 2010*, pp.177-181.

[11] Duke, J.; Friedlin, J.; Ryan, P. (2011) A Quantitative Analysis of Adverse Events and "Overwarning" in Drug Labeling, *Arch Intern Med* Vol.171, No.10, 2011, 944–946.

[12] Bisgin, H.; Liu, Z.; Fang, H.; Xu, X.;Tong, W.(2011) Mining FDA drug labels using an unsupervised learning technique - topic modeling, *BMC Bioinformatics*12(Suppl 10), S11.

[13] Richard, B.; Gregory, G.; Henk, H.(2012) Using Natural Language Processing to Extract Drug-Drug Interaction Information from Package Inserts, *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pp.206-213, Montréal, Canada, 2012, Association for Computational Linguistics.

[14] FDA (2008) Structured Product Labeling Resources http://www.fda.gov/ForIndustry/DataStandards/.

[15] Suzuki, S.; Koinuma, M.; Hidaka, Y.; Koike, K.; Nakamura, H.(2009) The Consciousness Research and Analysis on the Directive Pharmacists Who Provide Pre-education Prior to Clinical Practice?An Effort in the College of Pharmacy Nihon University?*YAKUGAKU ZASSHI* Vol. 129, No.9, 1103-1112, 2009

[16] Otani, Y.; Kawanaka, H.; Yoshikawa T.; Yamamoto, K.; Shinogi, T.; Tsuruoka S.(2005) Keyword Extraction from Incident Reports and Keyword Map Generation Method Using Self Organizing Map *Proceedings of IEEE International Conference on Systems, Man and Cybernetics 2005*, pp.1030–1035 .

[17] Kawanaka, H.; Otani, Y.; Yamamoto, K.; Shinogi, T.; Tsuruoka S.(2007) Tendency Discovery from Incident Report Map Generated by Self Organizing Map and its Development *Proceedings of IEEE International Conference on Systems, Man and Cybernetics 2007*, pp.2016–2021.

[18] Baba, T.; Liu, L.; Hirokawa S.(2010) Formal Concept Analysis of Medical Incident Reports *KES 2010*, Part III, LNAI 6278, pp. 207–214.

INTECH