MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010
© 2010 IMIA and SAHIA. All rights reserved.
doi:10.3233/978-1-60750-588-4-739

Extraction of Adverse Drug Effects from Clinical Records

Eiji Aramaki[a], Yasuhide Miura[b], Masatsugu Tonoike[b], Tomoko Ohkuma[b],
Hiroshi Masuichi[b], Kayo Waki[c], Kazuhiko Ohe[c]

[a] *Center for Knowledge Structuring, University of Tokyo, Japan*
[b] *Fuji Xerox, Japan*
[c] *University of Tokyo Hospital, Japan*

Abstract

With the rapidly growing use of electronic health records, the possibility of large-scale clinical information extraction has drawn much attention. We aim to extract adverse drug events and effects from records. As the first step of this challenge, this study assessed (1) how much adverse-effect information is contained in records, and (2) automatic extracting accuracy of the current standard Natural Language Processing (NLP) system. Results revealed that 7.7% of records include adverse event information, and that 59% of them (4.5% in total) can be extracted automatically. This result is particularly encouraging, considering the massive amounts of records, which are increasing daily.

Keywords:

Adverse effect, Side effect, Drug trial, Natural language processing (NLP)

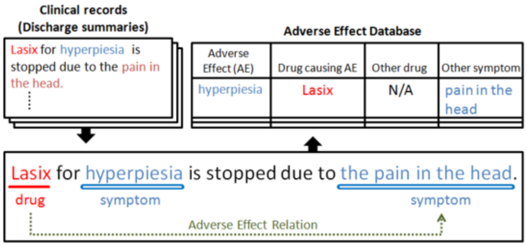
Introduction

The use of Electronic Health Records (EHR) in hospitals is increasing rapidly everywhere. They contain much clinical information about a patient's health, including the frequency of drug usage, related side-effects, and so on, which facilitates unprecedented large-scale research. Nevertheless, extracting clinical information from the reports is not easy because they are written in natural language. This study specifically examines adverse effects and event information buried in EHR.

Why is Adverse Effect Information needed?

For each approved drug, adverse effects are investigated through multiple phases of clinical trials. Clinical trials usually target only a single drug. Consequently, it is difficult to capture detailed effects resulting from multiple drug administration.

Real patients sometimes take multiple medications (e.g. prophylactic administration), leading to a gap separating the clinical trials and the actual use of drugs by patients. For ensuring patient safety, it is extremely important to bridge that gap.


Clinical records
(Discharge summaries)
Lasix for hyperpiesia is stopped due to the pain in the head.
Adverse Effect Database
Lasix for hyperpiesia is stopped due to the pain in the head.
drug
symptom
symptom
Adverse Effect Relation

Adverse Effect (AE)	Drug causing AE	Other drug	Other symptom
hyperpiesia	Lasix	N/A	pain in the head

Figure 1- Proposed approach

Adverse Effect Extraction

For such situations, we have started a project to extract adverse effect information from clinical records. As the first step, this paper presents examination of the following two questions.

(1) How much adverse information is included in records? To investigate this, we manually checked the information included in hundreds of records

(2) How to extract the information? Manual checking of all records takes too much time. Therefore, we attempted automatic extraction. We regard the adverse effect extraction task as including two sub-tasks (STEP 1) "term identification" and (STEP 2) "relation extraction."

STEP 1: Term identification: First, the system identifies drug and symptom expressions in records. This task is almost equivalent to the Named Entity Recognition (NER) task in NLP. We use a state-of-the-art NER method: conditional random fields (CRF) [1].

STEP 2: Relation identification: Then the system identifies which effect is related to which drug (adverse effect relation between a drug and an effect). We use both a pattern-based method and machine learning (SVM[2]) based method.

Through STEP 1 and STEP 2, a pair of a drug and an adverse effect is extracted along with other information (other drugs and symptoms) and is stored in a database (Fig. 1).

Although experiments described in this paper are related to Japanese medical reports, the proposed method does not depend on a specific language or domain.

Specific Research Questions

The specific goals of this study were the following:

1. To investigate how much adverse effect information exists in the clinical records (described in Section 2; *Materials*)

2. To investigate the accuracy with which the current technique (automatically) was able to extract adverse effect information (described in Section 4; *Experiments*)

Materials

This section introduces the materials (clinical records) used for this study, and reports summary data related to the adverse effects contained in the material.

Clinical Records (Discharge Summaries)

The materials of this study are 3,012 discharge summaries[1], which are reports generated by medical personnel at the end of a patient's hospital stay. The summaries were gathered from all departments of the University of Tokyo Hospital[2]. Because it costs much time to survey all summaries manually, we split the summaries into two sets: SET A, which contains keywords related to adverse effects (a keyword set is presented in Table 1); and SET B, which contains no keywords. Consequently, we obtained SET A consisting of 435 summaries, and SET B consisting of 2,577 summaries. Regarding SET A, we manually checked all of them. For SET B, four annotators checked small parts (randomly sampled) of them. Cases of ambiguity were resolved through discussion. We regarded even a suspicion of an adverse effect as positive data.

Quantities of Adverse Effects in Summaries

The results are presented in Fig. 2. For SET A, 53.5% (=233/435) of summaries described adverse effects. For SET B, 11.3% (=6/53) summaries described adverse effects. The ratio of SET A: SET B was 435:2577 (SET A=14.5%: SET B=85.5%). To sum up the results, we estimated that at least 7.7% (=0.145×0.535; only SET A) of summaries contain a description of adverse effects. Even considering that the result includes merely a suspicion of adverse effects, the summaries are a valuable resource for assessing adverse effects.

Annotation for Machine Learning (SET A Only)

To use a machine learning method, we also added tags to records. This annotation is limited to SET A because the other set (SET B) included few descriptions of adverse effects. The annotation includes information of two types: (1) term annotation, and (2) relation annotation.

(1) Term Annotation: Term annotation includes two tag types: (a) an expression for a drug, and (b) an expression for an effect. The definition is presented in Table 2. We annotated 1,045 drugs and 3,601 possible effects.

(2) Relation Annotation: Adverse effects were annotated. We represent the effect as a relation between a drug <d> and a symptom <s>, which is represented as a "relation" attribute. Table 3 shows several examples, wherein "relation =1" indicates an ID of a drug – adverse effect relation, which is a unique number in the text. We annotated 460 relations.

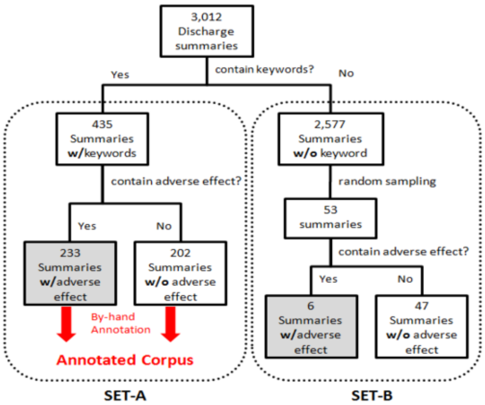

3,012 Discharge summaries
contain keywords?
Yes
No
435 Summaries w/keyword
contain adverse effect?
Yes
No
233 Summaries w/adverse effect
202 Summaries w/o adverse effect
By-hand Annotation
Annotated Corpus
2,577 Summaries w/o keyword
random sampling
53 summaries
contain adverse effect?
Yes
No
6 Summaries w/adverse effect
47 Summaries w/o adverse effect
SET-A
SET-B

Figure 2 - Data Configuration

Table 1 - Set of Keywords

Stop, stepped, Change changed, adverse effect, side effect

Table 2 - Markup Scheme (Tags and Definitions)

Tag	Definition (Examples)
s (symptom)	An expression of disease or symptom: e.g. *endometrial cancer, headache.* This tag covers not only a noun phrase but also a verb phrase such as "<s>*feel a pain in front of head*</s>".
d (drug)	An expression of medication of administration of a drug. Some examples are *Levofloxacin, Flexeril*

Table 3 - Annotation Example

<d relation="1">*ridora*</d>*resumed because it is associated with an* <s relation="1"> *eczematous rash* </s>
<d relation="1">*ACTOS(30)*</d> *brought both* <s relation="1">*headache*</s> *and* <s relation="1">*insomnia*</s>

* If a drug has two or more side effects, then they share the same ID. For example, *ACTOS* has two symptoms in the following:

Methods

The prior section explained that 7.7% of the records contain adverse-effect information. This section describes the standard two-step NLP used to extract information automatically.

[1] That amount roughly corresponds to summaries accumulated in one month at the University of Tokyo Hospital.

[2] All private information was removed from them. The definition of private information was referred from the HIPAA guidelines.

STEP 1: Term Identification

First, terms in records are identified. This task is similar to Named Entity Recognition (NER). Therefore, we use a state-of-the art NER method (conditional random fields (CRFs)), which has been shown to provide high performance for many tasks, such as part-of-speech tagging [1], text chunking [3], information extraction [4], and named entity recognition [5]. The detailed manner is described in a previous report [6].
In learning, we use standard parameters[3] and features as presented in Table 4. The only difference between the previous studies and this method is the dictionary feature.

STEP 2: Relation Identification

Then, the system decides which drug caused which symptom. For this identification, we compared two methods; a pattern based method, and a Support Vector Machine (SVM) based method.

Table 4 - Features for Term Identification

Lexicon & Stem	Target word (and its stem) and its surrounding words (and stem). The window size is five words (-2, -1, 0, 1, 2).
POS	Part of speech of TW and its surrounding words (-2, -1, 0, 1, 2). The part of speech is analyzed using a POS tagger[4].
DIC	A fragment for the target word appears in the medical dictionary MedDRA/J [7] consisting of covered 47,665 terms (lower level terms), and a drug name list (30,085 terms), which comes from drug package inserts [8].

Table 5 - Relation Identification Algorithm

```
procedure Relation_Identify ($D, S, K, n$)
for each drug $d$ in $D$ do
    for each symptom $s$ in $S$ do
        for each symptom $k$ in $K$ do
        pattern_based_identifier ($d, s, k, n$)
```

* D is a set of drugs in the target record; S is a set of symptoms in the record; K is a set of keywords shown in Table 1; n is the parameter of the pattern (window size)

Pattern based relation identifier: The procedure is presented in Table 5. The system judges whether each extracted term pair (d and s) has an adverse effect relation or not. The judgment is based on heuristic-rule-based patterns (Table 6).

Table 6 - Patterns for Relation Identification

*d*s*k*	*s*d*k*	*k*s*d*
*d*k* s*	*s*k*d*	*k*d*s*

"*" represents a wildcard for n words, where n is a parameter of window size

For example, given the example in Figure 3, the pattern "*d*k*s*" identifies an "*ACTOS-edema*" relation. Although the pattern is simple, it might suggest the difficulty of the task.

SVM based relation identifier: The SVM based method utilizes features as shown in Table 7 instead of patterns. The features come from words from a drug and a symptom. For example, we regard "*but stop for relief of the*" as a "word chain" feature in the Figure 3. For training, we regard a pair of a drug and a symptom sharing the same relation id as a positive sample, and the other pairs as negative samples. We utilized an RBF kernel, which has two parameters (C and gamma). We checked the performance with various parameter settings.

Table 7 - Features for Relation Identification

Symptom Lexicon	A symptom term
Drug Lexicon	A drug term
Word Chain	A series of words between a symptom and a drug.
Distance	A distance (the number of characters) between a drug term and a symptom term.

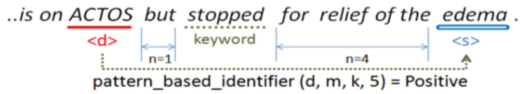

..is on ACTOS but stopped for relief of the edema .
<d>
keyword
<s>
n=1
n=4
pattern_based_identifier (d, m, k, 5) = Positive

Figure 3 - Relation Extraction Example

Experiments

We investigate performance of two types: (1) term identification, and (2) relation extraction. The experimental design is portrayed in Figure 4.

Experiment 1 (Term Identification)

Experimental Setting: We collected 435 Japanese discharge summaries, as described in Section 2.

Evaluation: We conducted experiments in a ten-fold cross validation manner. The performance is evaluated in the precision, recall, and *F*-measure attributable to the standard NE manner.

Results: Table 8 shows that we obtained all scores of more than 80%. This accuracy is higher than the Japanese NE task result (shown in IREX [10]) in which the best system's accuracy is *F*-measure of 0.68. That result indicates that the term identification in records is easier than the other tasks.

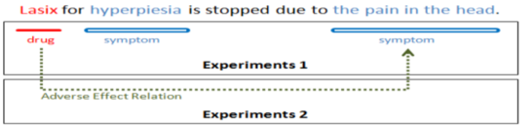


Figure 4 - Experimental Design

[3] f=3 c=1.5 window size=3

[4] http://chasen-legacy.sourceforge.jp/

Table 8 - Term Identification Results

	Precision	Recall	*F*-measure
s (Symptom)	0.855	0.802	0.828
d (Drug)	0.869	0.813	0.840

Experiment 2 (Relation Identification)

Experimental Setting: Because this experiment specifically examines relation identification performance, we adopt an oracle setting, wherein terms in a text are identified correctly.

Evaluation: The evaluation manner is identical to that in Experiment 1; ten-fold validation, precision, Recall and *F*-measure. We compared two methods: pattern-based (PTN) and SVM based (SVM). We checked the performance of various parameters. The F-measure curve in SVM is shown in Figure 5. We checked the possible combinations of two parameters and picked up the highest f-measure points (Table 9).

Results: Both PTN and SVM F-measures were lower than 0.65, indicating this is difficult task. Especially, SVM obtained a significantly lower performance than PTN (p=.05). One of the reasons is the amount of training data (especially positive data) is too small to capture the complex phenomena.

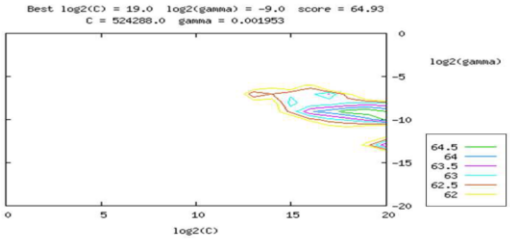


Figure 5 - F-measure in Various Parameters (C & gamma)

Table 9 - Relation Identification Results

	Precision	Recall	F-measure
PTN	0.411	0.917	0.650
SVM	0.576	0.623	0.598

Discussion

The experimental results revealed two salient facts.

1. How much information related to adverse effect is included in discharge summaries?

From the material, we infer that about 7.7% of the summaries contain information related to adverse effects.

2. To what extent are adverse effects extracted automatically?

The overall accuracy is estimated using the combined accuracies in experiment 1 and experiment 2: (accuracy of syndrome identification) × (that of drug) × (that of relation identification). Table 10 shows the result. Although the accuracy is insufficient (see precision 0.30), the proposed method (both SVM and PTN) could control the balance of precision and recall (Figure 6), which enables several practical applications appear promising: automatic mining under a high-precision setting, or pre-processing for human checks under the high-recall setting.

Table 10 - Results of Overall Accuracy

PTN	Precision	0.301 (=0.855 × 0.869 × 0.411)
	recall	0.597 (=0.802 × 0.813 × 0.917)

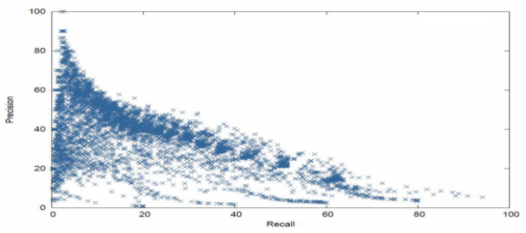


Figure 6 - Precision & Recall curve in SVM

Remaining problems

Training data: Compared with term identification, relation identification has low accuracy, which degrades the overall accuracy. Usually, the relation identification is solved using a machine learning approach (see a series of shared tasks [8]), we use that approach only slightly because adverse effects are rare events in records (the positive data are few).

Future studies should (1) increase training data to incorporate machine-learning techniques, or (2) apply another technique that works with small samples, such as active learning.

***Variants**:* Another problem is orthographic variants. A typical example is "*WBC decrease*", "*WBC depression*" or "*reduced WBC*" that share the same concept, but which have different expressions. Such variants engender serious problems in the symptom-aggregating process. In the future, normalizing techniques are highly desired.

Demonstration System

The presented system is available on the web (Figure 7). The annotation guidelines and sample corpora are also downloadable.

Figure 7 - Demonstration System

** free text input in the left window is converted into a database structure in the right window. http://luululu.com/text2table*

Related Work

Adverse Event/Effect Database

To date, several adverse effect databases are manually maintained, such as ARRS and GPRD.

The Adverse Event Reporting System (ARRS) is a famous adverse event database that is designed to support the FDA's safety program for all approved drugs. Reporting of adverse events from the point of care is voluntary in the United States. The current version of AERS contains 4,000,000 reports. The General Practice Research Database (GPRD) is a large database of medical records (over 3.6 million patients in the UK). Compared with the large databases described above, the data in this study are few and have low reliability. However, the automatic technique is highly desired considering the rapidly growing use of new medicines. We believe that the proposed automatic approach will be useful.

Related Natural Language Processing Studies

1. ***Term Identification***
Recent term identification uses machine-learning techniques such as Support Vector Machine (SVM) [2] and CRF [1]). Because of such trends, such techniques are also used in a clinical context [11,12]. We use the same approach as that in a previous study [11]; it effectively works with our corpus.

2. ***Relation Identification***
Relation identification has drawn much attention from various fields: Information Extraction (IE) fields such as MUC [13] and ACE [14], semantic relations (such as Semantic Relation tasks at SemEval2007 [15]) and Protein–Protein Interaction ontology. In all fields described above, machine-learning-based approaches using annotated corpora are popular.
This study deals with low-frequency phenomena (adverse events). Therefore, machine-learning approaches suffer from a paucity of positive data. As described in the *Discussion* section, we must solve this problem.

Conclusion

The method described in this paper extracts adverse drug events from texts in records. First, we annotated 435 discharge summaries with adverse effect information. Then, using the corpus, we investigated how much the current NLP system extracted the information. The results revealed that 7.7% of the records contain adverse event information; 59% of them (4.5% in all) were extracted (recall 59%; precision 30%). This result is encouraging, especially considering the massive and continually accumulating amounts of records. In the future, a high precision method is highly desired.

Acknowledgments

Part of this research is supported by a Grant-in-Aid for Scientific Research (A) of Japan Society for the Promotion of Science Project Number: 20680006 F.Y. 2008–2011 and a Research Collaboration Project with Fuji Xerox Co., Ltd.

References

[1] Lafferty J, McCallum A, and Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. Int Conf on Machine Learning 2001: 282–289.

[2] Vapnik VN. The Nature of Statistical Learning Theory. Springer, 1998.

[3] Sha F, and Pereira F. Shallow parsing with conditional random fields. Technical Report CIS TRMS-CIS-02-35, University of Pennsylvania; 2003.

[4] Pinto D, McCallum A, Lee X, and Croft W. Table Extraction using Conditional Random Fields. In: Proc. 26th ACM SIGIR, 2003:235–242.

[5] McCallum A and Li W. Early Results for Named Entity Recognition with Conditional Random Fields. In: Proc. 7th Conf on Natural Language Learning; 2003: 188–191.

[6] Aramaki E, Miura Y, Tonoike M, Ohkuma T, Mashuichi H, and Ohe K. TEXT2TABLE: Medical Text Summarization System Based on Named Entity Recognition and Modality Identification, Proc. HLT-NAACL2003 Workshop on BioNLP, 2009: 185–192.

[7] http://www.meddramsso.com/MSSOWeb/

[8] http://www.info.pmda.go.jp/

[9] http://biocreative.sourceforge.net/

[10] Sekine S andEriguchi Y, Japanese named entity extraction evaluation: analysis of results. Proc. 18th Conf on Computational Linguistics, 2000: 1106–1110.

[11] Aramaki E, Imai T, Miyo K and Ohe K. Automatic Deidentification by using Sentence Features and Label Consistency, Workshop on Challenges in Natural Language Processing for Clinical Data, 2006.

[12] Tawanda S and Ozlem U. Role of Local Context in Automatic Deidentification of Ungrammatical, Fragmented Text. In: Proc. HLT-NAACL2006; 2006: 65–73.

[13] Grishman R and Sundheim B. Message understanding conference – 6: A brief history. In Proc. Int Conf on Computational Linguistics, 1996: 466–471.

[14] http://www.itl.nist.gov/iaui/894.01/tests/ace/

[15] Girju R, Szpakowicz S, Nakov P, Turney P, Nastase V and Yuret D. SemEval-2007 task 04: Classification of semantic relations between nominals, ACL2007 Workshop on Semantic Evaluations 2007: 464–467.

Address for correspondence

Eiji ARAMAKI <eiji.aramaki@gmail.com> Center for Knowledge Structuring, University of Tokyo. Tokyo 113-8655, Japan.