2004−ICS−138 (24)
2004／12／5

Information Extraction from MEDLINE abstracts of clinical trials

Kazuo Hara† and Yuji Matsumoto†

In this paper, we report experiment results on applying IE methodology to extract "compared treatments", "primary endpoints" and "patient population" from MEDLINE abstracts of clinical trials.

1. Introduction

Recently, people engaged in medical treatment have been paying more attention on the notion of Evidence-Based Medicine (EBM), that is to say, "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients"[1]. In practicing EBM, they are encouraged to be well posted on up-to-date sources of medical knowledge such as MEDLINE[2], the US National Library of Medicine's bibliographic database covering the fields of medical, pharmaceutical and biological sciences. Among MEDLINE abstracts, those about clinical trials play one of the most important roles in the EBM, for the results of clinical trials can provide firm evidence for applying therapy in actual medical treatments. However, the rate at which new articles are being introduced into the MEDLINE database is fairly high, it takes patients or doctors who seek beneficial knowledge quite some time to read all of the articles which may contain clues of finding suitable therapy. So, in order to support members of the medical community, we extract summary information from MEDLINE abstracts of clinical trials, in an effort to reduce the amount of time required to find relevant medical information.

In the research field of natural language processing, the task of information extraction (IE) has been pursued with a great deal of interests for decades. For example, in the series of Message Understanding Conferences (MUC), participants developed methods for extracting information of the scenario-templates presented by the conference organizer[3]. The focus of the study there was the construction of domain-specific lexicons and extraction patterns based on human judgment. Following MUC, people's attention has shifted to automatic knowledge acquisition including lexicons and patterns. In this paper, however, we use conventional IE methods to conduct a preliminary experiment in extraction from MEDLINE abstracts of clinical trials.

We describe the information extracted from the MEDLINE abstracts in section 2. We show the extraction methods in section 3, describe the results of the experiment in section 4, and finally conclude with some discussion and perspective on future work in section 5.

2. IE targets from the abstracts of clinical trials

In general, clinical trials should be carried out on pre-determined conditions set for "compared treatments", "primary endpoints", "patient population" and so on, to minimize bias and maximize precision of the results of the analysis[4]. In other words, we can at least say that the above three should be the target information to extract because they best characterize the clinical trials. Examples are shown in Table 2.

2.1 Compared treatments

The final goal of clinical trials is to find out whether the investigational therapy has clinical benefits. For example, in most confirmatory trials, the candidate product for a new drug is compared with control treatments or active comparators. In this paper, "compared treatments" correspond to the drug or therapy compared in the clinical trials.

2.2 Primary endpoints

In general, clinical trials investigate the efficacy and safety of a new drug or therapy. As a primary endpoint, the variable directly related to the patient's outcome, such as decrease in mortality or improvement in quality of life, is

† Nara Institute of Science and Technology

often used in confirmatory trials. In contrast, indirect variables that include pharmacokinetics parameters are used in many exploratory trials.

2.3 Patient population

The inclusion criteria for patients is set up in clinical trials. For example, only healthy male adults are enrolled in phase I trials in many cases.

3. Method

3.1 POS tagging and NP chunking

After tokenization, word sequences in the MEDLINE abstracts are part-of-speech (POS) tagged using the TnT tagger[5]. Then, noun phrase (NP) chunking is performed by the YamCha chunker[6]. The NP chunking process is illustrated in Figure 1.

3.2 Classification of chunked NP

NPs are manually classified into 11 groups according to the following tag sets.

DISEASE: for the name of disease or virus.
DRUG: for chemical compounds or drugs.
STUDY: for clinical trial or statistical analysis.
THERAPY: for general expression of treatments.
PATIENT: for participants in the clinical trials.
TARGET: for endpoints or clinical laboratory evaluation.
SCHEDULE: for time schedule in clinical trials.
VALUE: for expression of variation corresponding to TARGET.
NUMBER: for numerals.

Table 1 The classification tags for the chunked NPs and their corresponding examples.

tag	example
DISEASE	"chronic hepatitis C"
DRUG	"interferon plus ribavirin"
STUDY	"clinical trial"
THERAPY	"antiviral treatment"
PATIENT	"HBeAg-positive patients"
TARGET	"efficacy and safety"
SCHEDULE	"an additional 24 weeks"
VALUE	"significantly higher rates"
NUMBER	"20 percent"
OTHERS	"we"
FALSE_TAG	(for chunking error)

OTHERS: for NPs which are not classified into the above groups.
FALSE_TAG: for NP chunking error.

Notice that "SCHEDULE", "VALUE", and "NUMBER" may be consolidated into one tag because we don't use those tags in the IE rules described in the next section. The tags and their examples are listed in Table 1.

3.3 Information extraction rules

The most of MEDLINE abstracts consist of a title and a main text. Titles are expected to contain important information, so we set two rules for titles and main texts separately.

3.3.1 Rule 1 (for titles)

compared treatments:
(1) NPs classified into "[DRUG]", and
(2) NPs classified into "[THERAPY]"
appeared in titles are extracted.

primary endpoints:
(1) NPs classified into "[TARGET]"
appeared in titles are extracted.

patient population:
(1) NPs classified into "[PATIENT]", and
(2) phrases matched with the following pattern for patient restricted by disease: "[PATIENT] with [DISEASE]", and
(3) phrases matched with the following pattern for patient restricted by other clinical conditions: "[PATIENT] with [TARGET]"
appeared in titles are extracted.

3.3.2 Rule 2 (for main texts)

All phrases which matched the following patterns containing regular expressions appeared in main texts are extracted.

compared treatments:
(1) pattern for listing drugs: "[DRUG] (.* as [non-NP] | .* as [OTHERS]) [DRUG]"
(2) pattern for comparing drugs or therapy: "("Compar*" | "compar*" | "between") * ([DRUG] | [THERAPY]) * (" versus " | "with" | "and") * ([DRUG] | [THERAPY])"

primary endpoints:
(1) pattern for what authors performed: "("We" | "we" as [OTHERS]) .* [TARGET]"

Table 2 Targets for information extraction: from the 3 abstracts out of 50 used in the experiment.

PMID	compared treatments	primary endpoints	patient population
15282352	peginterferon alfa-2a plus ribavirin, interferon alfa-2a plus ribavirin	sustained virologic response	patients with chronic HCV infection and HIV
15235875	iron reduction therapy, regular blood tests	serum alanine aminotransferase levels	patients with CHC
15131791	epoetin alfa, placebo	quality of life, increase hemoglobin	HCV-infected patients

Original sentence:

- We conducted a multicenter, randomized trial comparing peginterferon plus ribavirin with interferon plus ribavirin for the treatment of chronic hepatitis C in persons coinfected with HIV.

NP-chunked sentence:

- [We] conducted [a multicenter, randomized trial] comparing [peginterferon plus ribavirin] with [interferon plus ribavirin] for [the treatment] of [chronic hepatitis C]in [persons] coinfected with [HIV].

NP-tagged sentence:

- [GENERAL] conducted [STUDY] comparing [DRUG] with [DRUG] for [THERAPY] of [DISEASE] in [PATIENT] coinfected with [DISEASE].

Patterns matched with rule 2:

- compared treatments: comparing [DRUG] with [DRUG]
- patient population: [PATIENT] * with [DISEASE].

Extracted information:

- compared treatments: peginterferon plus ribavirin, interferon plus ribavirin.
- patient population: persons coinfected with HIV.

Fig. 1 An example of the process of NP chunking (chunked NPs are enclosed in the brackets) and an example of extracted information.

(2) pattern for what performed in the clinical trials: "("This*" | "this*" as [STUDY]) .* [TARGET]"

patient population:

(1) pattern for patient restricted by disease: "[PATIENT] * with [DISEASE]"

(2) pattern for patient restricted by other clinical conditions: "[PATIENT] * with [TARGET]"

In Figure 1, an example of information extraction using rule 2 is shown.

4. Experiment

4.1 Data

We downloaded 50 most recent abstracts of clinical trials from the MEDLINE database. The important part of reference query was "hepatitis[MeSH Terms] AND hasabstract[text] AND Randomized Controlled Trial[ptyp]". To simplify the experiment, abstracts were selected from the medical area of hepatitis. Details of the 50 abstracts used in the experiment are shown in Table 6.

4.2 Results

The results from rule 1 applied for titles alone and rule 1 plus rule 2 applied for all texts in abstracts are summarized in Table 4. For each IE target, there are cases where multiple objects are to be extracted from abstracts, so we divided recall into two types; recall (a): all information was extracted, recall (b): at least a part of information was extracted.

The results from rule 1 for titles can be considered as the baseline, because just putting together the titles is close to summarizing the articles. Judging from the results of experiments, although the results of IE from all texts are su-

Table 3 The distribution of NPs from the 50 abstracts used in the experiment.

tag	number	percent
DISEASE	292	8%
DRUG	341	10%
STUDY	135	4%
THERAPY	267	8%
PATIENT	501	14%
TARGET	602	17%
SCHEDULE	208	6%
VALUE	303	9%
NUMBER	205	6%
OTHERS	507	15%
FALSE_TAG	112	3%
total	3473	100%

Table 4 The results of information extraction from the 50 abstracts used in the experiment.
recall (a): all information was extracted.
recall (b): at least a part of information was extracted.

rule	IE target	precision	recall (a)	recall (b)
rule1	compared treatments	82%	40%	92%
	primary endpoints	73%	24%	32%
	patient population	94%	48%	60%
rule1 + rule2	compared treatments	85%	64%	92%
	primary endpoints	77%	52%	64%
	patient population	76%	82%	86%

perior to those from titles alone, the rules we set in the experiment are not powerful enough for practical use. Another remark is that the lower recall in rule 1 implies that few authors seem to describe primary endpoints in the title.

In addition, we evaluated the result of NP chunking and the reproducibility of manual classification of the NPs for the tag sets. Table 3 shows the distribution of NPs from the 50 abstracts used in the experiment. The error rate of NP chunking is not so large (3%). Table 5 shows the result of 5-fold cross validation test for automatic classification of the NPs using YamCha again; of course, the learning model is different from the one for NP chunking in section 3.1. Here, YamCha classifies NPs using 6 features; the class of the phrase (such as NP, VP and PP), the first and the last word of the phrase, the first and the last POS of the phrase, and the phrase itself. For each NP, YamCha calculates the most probable tag using the features of the words or phrases around the target NP. We can say that tags with higher precision and recall such as "PATIENT" or "NUMBER" are comparatively correct. However, other tags seem to have some points which should be improved.

Table 5 The result of 5-fold cross validation test for automatic classification of the NPs using YamCha.

tag	precision	recall
DISEASE	79%	63%
DRUG	64%	62%
STUDY	74%	46%
THERAPY	89%	64%
PATIENT	88%	85%
TARGET	54%	71%
SCHEDULE	83%	68%
VALUE	62%	61%
NUMBER	78%	75%
OTHERS	52%	66%
FALSE_TAG	67%	14%

5. Discussion

We have described a preliminary experiment in the task of information extraction from MEDLINE abstracts of clinical trials. The problems that developed are: (1) improvement

Table 6 Overview of the 50 abstracts used in the experiments: published journal, type of clinical trial and PMID.

index	details of the 50 abstracts
published journal	5 abstracts: J Viral Hepat. 4 abstracts: Hepatology. 3 abstracts: Aliment Pharmacol Ther., Gastroenterology., Hepatobiliary Pancreat Dis Int., N Engl J Med.
type of clinical trial	27 abstracts: confirmatory trial 9 abstracts: exploratory trial 9 abstracts: cohort study or case-control study 4 abstracts: reusing data of past clinical trials 1 abstracts: other study
PMID	15371578, 15349912, 15334776, 15330905, 15282352, 15282351, 15276594, 15264116, 15235875, 15233669, 15226169, 15206691, 15192274, 15188169, 15166532, 15163089, 15147122, 15139984, 15138107, 15131791, 15131468, 15125329, 15122768, 15122749, 15117326, 15115965, 15108656, 15059069, 15057903, 15057741, 15040533, 15031786, 14999598, 14996676, 14996351, 14987322, 14986815, 14984381, 14984379, 14980101, 14976882, 14969841, 14969836, 14745327, 14743544, 14738561, 14738560, 14738559, 14733779, 14733608

of rules for IE and (2) elaboration of tag sets for NPs.

As concerning rules for IE, our rules used in the experiment are based on heuristics. There is disadvantage in heuristic methods. First of all, the human ability to describe patterns is not always complete. Rewriting rules whenever inconsistency is found has much greater costs in the long run. Furthermore, even the rules made out by experts have no theoretical guarantee that they are correct. So we are trying to consider comprehensive and exhaustive methods, such as that proposed in Sudo et al., 2003[7]. For the elaboration of tag sets, MeSH, the National Library of Medicine's controlled vocabulary thesaurus[8], will be helpful at least for "DISEASE" and "DRUG."

Two of the topics for future work are (3) expansion of the range of applicable abstracts and (4) augmentation of IE targets. We dealt with abstracts of clinical trials only related to hepatitis in our experiments. Expanding to other medical treatments may affect the rules and the tag sets. The augmentation of IE targets will include the number of patients or significance of superiority which make up the evidence for the therapy.

References

1) Sackett, D.L. et al. (1996) Evidence based medicine: what it is and what it isn't. BMJ 312 (7023), 13 January, 71-72.
2) MEDLINE. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
3) Overview of MUC-7/MET-2. (1998) http://www.muc.saic.com/proceedings/muc_7_proceedings/overview.html
4) International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use. (1998) Statistical Principles for Clinical Trials. (ICH E9)
5) Thorsten Brants. (2000) TnT - a statistical part-of-speech tagger, ANLP 2000.
6) Taku Kudo and Yuji Matsumoto. (2001) Chunking with Support Vector Machines, NAACL 2001.
7) Kiyoshi Sudo and Satoshi Sekine and Ralph Grishman. (2003) An Improved Extraction Pattern Representation Model for Automatic IE Pattern Acquisition, ACL 2003.
8) MeSH: Medical Subject Headings. http://www.nlm.nih.gov/mesh/meshhome.html