Information Extraction and Sentence Classification applied on Clinical Trial MEDLINE Abstracts

Nara Institute of Science and Technology
Kazuo Hara and Yuji Matsumoto

Contents

1. Backgrounds
2. Methods
   - ① Information Extraction
     by manually written regular expressions
   - ② Sentence Classification
     by automatically produced patterns by BACT
3. Experiments
4. Future Work

Backgrounds

- Needs in the medical society
  - Practice of Evidence Based Medicine
- MEDLINE
  - The US National Library of Medicine's bibliographic database including pharmaceutical domain
  - The most popular information source for finding evidence of new therapy
- Our Research Goal:
  - To construct the information extraction (IE) system from MEDLINE clinical trial abstracts

Targets of Information Extraction

- 「Compared Treatment」
  - The aim of clinical trial is to investigate the efficacy and safety of the new treatment comparing with current therapy.
- 「Endpoint」
  - To what the new treatment shows greatness?
- 「Patient population」
  - To whom the new treatment shows greatness?

Example of the Targets

- INPUT
  - 「We compared drug X and drug Y by investigating blood pressure and survival time in patients with high blood pressure.」
- OUTPUT
  - 「Compared Treatment」(two targets) → drug X, drug Y
  - 「Endpoint」(two targets) → blood pressure, survival time
  - 「Patient Population」(one target) → patients with high blood pressure

※ Take notice that each clinical trail has sometimes more than one IE target.

Research Goal

- To construct the information extraction (IE) system from MEDLINE clinical trial abstracts

【a MEDLINE abstract】

•TITLE: **Peginterferon Alfa-2a plus ribavirin** versus **interferon alfa-2a plus ribavirin** for chronic hepatitis C in HIV-coinfected persons.

•BACKGROUND: Chronic hepatitis C virus (HCV) infection is a cause of major … interferon plus ribavirin for the treatment of chronic hepatitis C in **persons coinfected with HIV**.

•METHODS: A total of 66 subjects were randomly assigned to receive … either a virologic response or histologic improvement.

•RESULTS: Treatment with peginterferon and ribavirin was associated with a significantly higher rate of **sustained virologic response** than was treatment with interferon and ribavirin. ・・・

【IE result】

(1) Compared Treatment:
・peginterferon alfa-2a plus ribavirin
・interferon alfa-2a plus ribavirin

(2) Endpoint:
・sustained virologic response

(3) Patient Population:
・persons coinfected with HIV

Contents

1. Backgrounds
2. Methods
   - ① Information Extraction
     by manually written regular expressions
   - ② Sentence Classification
     by automatically produced patterns by BACT
3. Experiments
4. Future Work

The preliminary process

INPUT sentence: from a clinical trial MEDLINE abstract

"We conducted a multi-center, randomized trial comparing peginterferon plus ribavirin with interferon plus ribavirin for the treatment of chronic hepatitis C in persons co-infected with HIV."

⇩

NP chunking: by YamCha (Kudo and Matsumoto, 2001)

"[We] conducted [a multi-center, randomized trial] comparing [peginterferon plus ribavirin] with [interferon plus ribavirin] for [the treatment] of [chronic hepatitis C] in [persons] co-infected with [HIV]."

⇩

NP tagging: manually by domain specific knowledge

"[We] conducted [STUDY] comparing [DRUG] with [DRUG] for [THERAPY] of [DISEASE] in [PATIENT] co-infected with [DISEASE]."

Tag definition

【Tag】	【covered concept】	【example】
DISEASE:	disease, symptom, virus	chronic hepatitis C
DRUG:	drug, chemical compound	interferon
STUDY:	clinical trial	clinical trial
THERAPY:	treatment, regimen	antiviral treatment
PATIENT:	participants in the trial	HBeAg-positive patients
TARGET:	endpoints	efficacy and safety
SCHEDULE:	time schedule of the trial	an additional 24 weeks
VALUE:	value of TARGET	significantly higher rates
NUMBER:	numeral expression	20 percent

Task definition

NP tagging: manually by domain specific knowledge

"[We] conducted [STUDY] comparing [DRUG] with [DRUG] for [THERAPY] of [DISEASE] in [PATIENT] co-infected with [DISEASE]."

② Sentence Classification: by BACT (**a Boosting Algorithm for Classification of Trees;** Kudo and Matsumoto, 2004)

⇨

Result:

Compared Treatment: +1 (Yes)

Endpoint: -1 (No)

Patient Population: +1 (Yes)

① Information Extraction:

Pattern matching by regular expressions

⇨

Result:

Compared Treatment : "peginterferon plus ribavirin"

Compared Treatment : "interferon plus ribavirin"

Endpoint : (none)

Patient Population : "persons co-infected with

Contents

1. Backgrounds
2. Methods
   - ① Information Extraction
     by manually written regular expressions
   - ② Sentence Classification
     by automatically produced patterns by BACT
3. Experiments
4. Future Work

① Information Extraction by regular expressions

NP tagging: manually by domain specific knowledge

"[We] conducted [STUDY] comparing **[DRUG]** with **[DRUG]** for [THERAPY] of [DISEASE] in **[PATIENT] co-infected with [DISEASE]**."

⇩

① Information Extraction: by manually written regular expressions

「Compared Treatment」: "compar∗" [**DRUG**] "with" [**DRUG**]

「Patient Population」: **[PATIENT] ∗ with [DISEASE]**

⇩

Result:

Compared Treatment : **"peginterferon plus ribavirin"**

Compared Treatment : **"interferon plus ribavirin"**

Endpoint : (none)

Patient Population : **"persons co-infected with HIV"**

② Sentence Classification

NP tagging: manually by domain specific knowledge

"[We] conducted [STUDY] comparing [DRUG] with [DRUG] for [THERAPY] of [DISEASE] in [PATIENT] co-infected with [DISEASE]."

⇩

② Sentence Classification: by BACT (Kudo and Matsumoto, 2004) , that automatically produce optimal patterns by machine learning.

⇩

Result:

Compared Treatment: +1 (Yes)

Endpoint: -1 (No)

Patient Population: +1 (Yes)

※ The accuracy of IE may improve if we apply ① Information Extraction to "yes" sentences only.

Example of Sentence Classification

"[We] conducted [STUDY] comparing [DRUG] with [DRUG] for [THERAPY] of [DISEASE] in [PATIENT] co-infected with [DISEASE]."

constructed patterns	Compared Treatment	Endpoint	Patient Population
(default)	-0.079	-0.141	-0.210
"We"	0.051	0.016	0.105
"STUDY"	0.013	0.065	0.081
"DRUG"	0.045	0.009	-0.003
"with"	0.008	-0.002	0.037
"with DRUG"	-0.003	–	-0.050
"for"	–	–	-0.006
"THERAPY"	0.014	–	-0.001
"for THERAPY"	-0.006	–	–
"of"	0.002	–	–
"DISEASE"	0.000	–	0.034
"of DISEASE"	-0.009	–	–
"in"	-0.013	0.012	0.000
"PATIENT"	0.007	-0.028	0.070
"in PATIENT"	–	0.000	–
"with DISEASE"	0.006	0.005	0.018
Total weight	0.035	-0.065	0.074
yes or no	**+ 1 (yes)**	**-1 (no)**	**+ 1 (yes)**

※ BACT calculates the weight for each patterns by learning from training data. (red numerals are minus weight)

Interpretation of patterns by BACT

automatically constructed patterns by BACT that include "DRUG"	Compared Treatment	Endpoint	Patient Population
"PATIENT received DRUG"	0.048	–	–
"DRUG"	0.046	–	–
"TARGET of DRUG"	–	0.035	–
"DRUG, DRUG"	0.013	–	–
"received DRUG"	0.01	0.023	–
"of DRUG"	0.006	0.012	–
"with DRUG"	-0.004	–	-0.026
"to DRUG"	-0.013	–	-0.012
"in DRUG"	-0.019	–	–

※ This table shows that not all NPs tagged by "DRUG" correspond to IE target of "Compared Treatment".

Contents

1. Backgrounds
2. Methods
   - ① Information Extraction
     by manually written regular expressions
   - ② Sentence Classification
     by automatically produced patterns by BACT
3. Experiments
4. Future Work

Data

We downloaded the 50 most recent abstracts of clinical trials from the MEDLINE database: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi on October 2004.

Query:

「"hepatitis"[MeSH Terms] AND hasabstract[text] AND Randomized Controlled Trial[ptyp]」

※ To simplify the experiment, abstracts were selected from the medical area of hepatitis.

Results

- Precision, Recall (5-fold cross validation)

	Compared Treatment		Endpoint		Patient Population	
	precision	recall	precision	recall	precision	recall
① Information Extraction	84.8%	64.0%	77.0%	52.0%	76.2%	82.0%
② Sentence Classification (dep)	86.8%	78.5%	84.7%	72.2%	75.2%	71.4%
② Sentence Classification (Ngram)	82.6%	71.7%	85.7%	73.2%	81.5%	81.5%

- dep is outperformed by Ngram with respect to "Patient Population". We can guess the reason here: parse errors occurred in many of the dependency trees caused by PP attachment ambiguity ("PATIENT with DISEASE").

Contents

1. Backgrounds
2. Methods
   - ① Information Extraction
     by manually written regular expressions
   - ② Sentence Classification
     by automatically produced patterns by BACT
3. Experiments
4. Future Work

Future Work

- Automatic NP Tagging
  - In this experiment, manually tagged by domain specific knowledge.
- Bigger Corpus
  - In this experiment, only 50 abstracts.
- Apply 「① Information Extraction」 to the result from 「② Sentence Classification」
  - How Sentence Classification contribute to the Information Extraction?