Problems associated with applying NLP techniques to Clinical Trial MEDLINE Abstracts

Computational Linguistics Laboratory
Kazuo Hara

1

Background & Aim

2

Ubiquitous Medicine - a trend in the medical community -

- This trend is supported by popularization of ubiquitous technology such as
  - Remote Diagnostic Imaging, and
  - Electronic Health Records.
- The community is going to share comparable clinical information among medical sites.

3

This trend leads to a demand for high quality medical treatments.

- The concept, Evidence-Based Medicine (EBM), has become prevalent recently.
  - EBM requires medical practitioners to select appropriate treatments for individual patients based on the current best evidence.
- Where does the current best evidence come from?
  - One major source of evidence is clinical trial results.

4

What are the clinical trials?

- Phase I
  - Examination of the safety of the new treatment.
- Phase II
  - Exploration of the usage and dosage of the new treatment.
- Phase III
  - Verification of the new treatment compared to an active control or placebo.
- Phase IV
  - Post Marketing Surveillance of the new treatment.

5

Where to access the clinical trial results information?

- MEDLINE, the U.S. National Library of Medicine's (NLM) database of biomedical citations and abstracts that is searchable on the Web.
- MEDLINE search index includes:
  - clinical trial phases (phase I, II, III, and IV),
- but does not include important keys such as:
  - "compared treatments", "patient population", and "endpoints".

6

A clinical trial result is always summarized in a table.

- A typical example (phase III)

	Treatment A (New Drug)	Treatment B (Active Control)	statistical significance
Endpoint (Efficacy)	value or score	value or score	p-value
Endpoint (Safety)	frequency or count	frequency or count	p-value

MEDLINE abstracts are just the rewriting of the result tables.

【 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. ・・・

【 Important Keys 】

(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**

Our research goal is:

- Extracting information with respect to important keys from each clinical trial MEDLINE abstract in order to construct a database which is easy to access.
  - The keys are:
    - "compared treatments", "patient population", and "endpoints".
- This can become a support for realizing EBM in the medical community.

Our previous approach

Text mining based on phrase-structure trees

Our previous approach consists of:

- Converting MEDLINE texts into phrase-structure trees using an NLP parser, and
- Mining these trees for patterns to find target information such as "compared treatments".

Resources

- NLP parser
  - Charniak's phrase-structure analyzer (Charniak, 2000)
- text miner
  - The sentence classifier or semi-structured text classifier proposed in (Kudo and Matsumoto, 2004)

Our previous approach

Pattern mining (an example)

Input Text: "We conducted STUDY comparing DRUG with DRUG for THERAPY of DISEASE in PATIENT co-infected with DISEASE."

Output: patterns for finding targets such as "Compared Treatment", and their weights

patterns in parsed phrase-structure trees	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
:	:	:	:
"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
Classification	**+ 1 (yes)**	**-1 (no)**	**+ 1 (yes)**

Our previous approach

However …

- There is a problem with applying NLP parsing techniques to MEDLINE abstracts.
  - Most NLP parsers have difficulty analyzing coordinate structures and prepositional phrases correctly.
  - Unknown technical terms also reduce the quality of parsing output.

14

Our previous approach

Coordinate structures

- " in 118 (80%) of the 148 evaluable patients in the standard arm "
- " in 129 (88%) of the 147 evaluable patients in the dose-dense arm "
  - These coordinate structures appear frequently in clinical trial MEDLINE abstracts.
  - These are likely to include important information about the clinical trial's design.

15

Our previous approach

To make parsing successful,

- Manually annotated MEDLINE corpus constructed by human labor is necessary, but is high cost.
- So, in addition to this approach, we plan another one.

16

Our ongoing approach

17

Our ongoing approach

Focus on the alignment of the coordinate structures

- Coordinate structures are likely to include important information about the clinical trial's design.
  - " in 118 (80%) of the 148 evaluable patients in the standard arm "
  - " in 129 (88%) of the 147 evaluable patients in the dose-dense arm "

18

Our ongoing approach

How to find and extract coordinate structures?

- (Kurohashi and Nagao, 1994):
- "A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures.“
  - Determine similarities (or weights) between tokens based on syntactic and semantic knowledge.
  - Calculate the similarity score between two token sequences according to their component token similarities.
  - A high similarity score indicates that the two token sequences construct coordinate structures.

19

Our ongoing approach

The concept in (Kurohashi and Nagao, 1994):

	in	118	(80%)	of	the	148	evaluable	patients	in	the	standard	arm
in	○	–	–	△	–	–	–	–	○	–	–	–
129	–	△	–	–	–	△	–	–	–	–	–	–
(88%)	–	–	△	–	–	–	–	–	–	–	–	–
of	△	–	–	○	–	–	–	–	△	–	–	–
the	–	–	–	–	○	–	–	–	–	○	–	–
147	–	△	–	–	–	△	–	–	–	–	–	–
evaluable	–	–	–	–	–	–	○	–	–	–	△	–
patients	–	–	–	–	–	–	–	○	–	–	–	△
in	○	–	–	△	–	–	–	–	○	–	–	–
the	–	–	–	–	○	–	–	–	–	○	–	–
dose-dense	–	–	–	–	–	–	△	–	–	–	△	–
arm	–	–	–	–	–	–	–	△	–	–	–	○

Our ongoing approach

Shortcomings of (Kurohashi and Nagao, 1994)

- Revolutionary for incorporating both syntactic and semantic similarity in identifying coordinate structures.
- However, ad-hoc token weightings may reduce accuracy to find coordination depending on the domain of texts.

21

Our ongoing approach

Improving (Kurohashi and Nagao, 1994)

- Develop a method that can learn similarities (weights) from the MEDLINE corpus using machine learning.
- Seed the vector used to identify coordinate structures with weights from similarity as measured with the above method.

22

Summary

- Background:
  - Ubiquitous medicine leads to a demand for high quality medical treatments represented by EBM.
- Our research goal is:
  - Extracting important information from clinical trial MEDLINE abstracts in order to support the realization of EBM.
- Our ongoing approach is:
  - Focusing on the coordinate structures and developing a method that can learn from a corpus using machine learning.

23