Extracting Clinical Trial Information from MEDLINE Abstracts

2006. 5. 25
Computational Linguistics Laboratory
Kazuo Hara

Background and Objectives

Ubiquitous Medicine
- a trend in the medical community -

- The 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.

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 phase III clinical trial results.

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.

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 has search indexes that include:
  - clinical trial phases (phase I, II, III, and IV),
- but does not include important keys such as:
  - "compared treatments" and "patient population“.

A clinical trial result is always summarized in tables according to keys.

- A typical example of a result table (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.

Keys in a clinical trial:

- Compared Treatments :
  - docetaxel
  - fluorouracil
- Patient Population :
  - patients with cancer

Corresponding expression in the MEDLINE abstract:

- “Phase III study comparing docetaxel with fluorouracil in patients with cancer …”

Our research goal is:

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

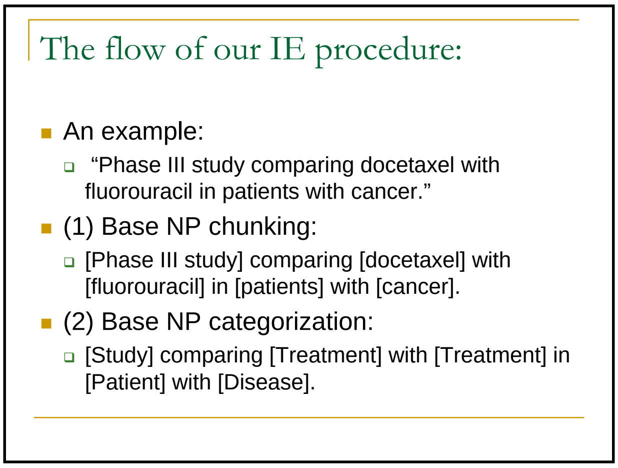
Purpose of today’s presentation

- To report results of experiment in extracting important information for EBM from the abstracts of phase III clinical trials,
  - in an effort to investigate how far the existing natural language processing (NLP) techniques could support EBM using MEDLINE database.

Information Extraction (IE) techniques applied to phase III abstracts

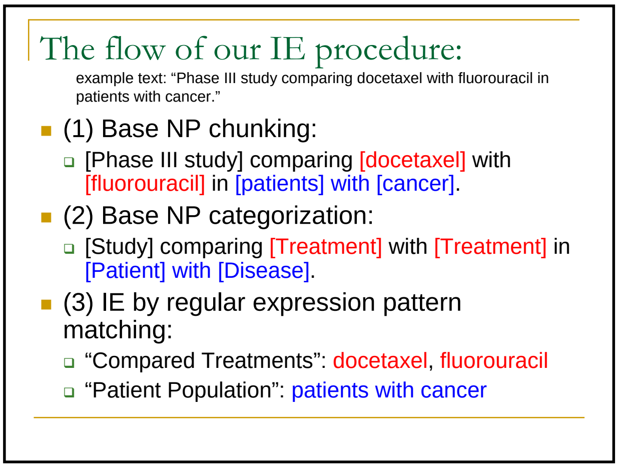
We use conventional IE techniques.

- (0) Part-of-speech tagging
  - TnT tagger (Brants, 2000)
- (1) Base Noun Phrase chunking
  - SVM based chunker: YamCha (Kudo and Matsumoto, 2001)
- (2) Base Noun Phrase categorization
  - SVM based categorizer: YamCha (Kudo and Matsumoto, 2001)
- (3) Information Extraction by regular expression pattern matching



The flow of our IE procedure:

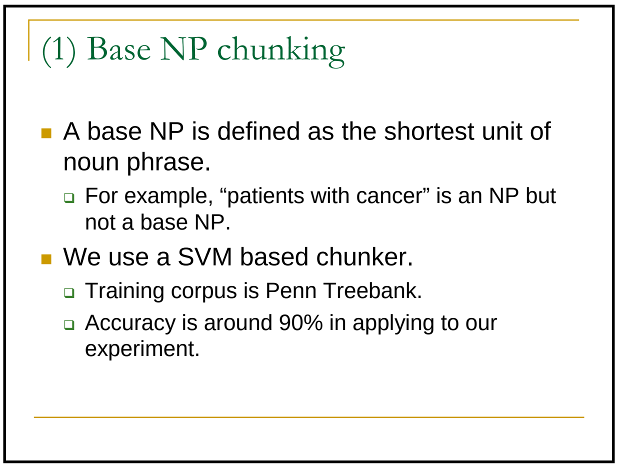
- An example:
  - “Phase III study comparing docetaxel with fluorouracil in patients with cancer.”
- (1) Base NP chunking:
  - [Phase III study] comparing [docetaxel] with [fluorouracil] in [patients] with [cancer].
- (2) Base NP categorization:
  - [Study] comparing [Treatment] with [Treatment] in [Patient] with [Disease].



The flow of our IE procedure:

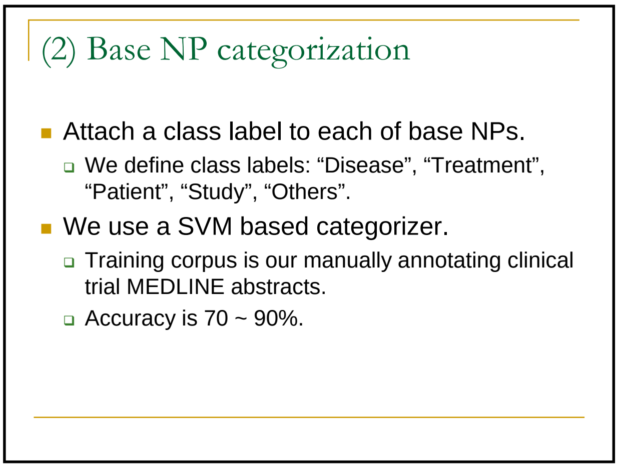
example text: “Phase III study comparing docetaxel with fluorouracil in patients with cancer.”

- (1) Base NP chunking:
  - [Phase III study] comparing [docetaxel] with [fluorouracil] in [patients] with [cancer].
- (2) Base NP categorization:
  - [Study] comparing [Treatment] with [Treatment] in [Patient] with [Disease].
- (3) IE by regular expression pattern matching:
  - “Compared Treatments”: docetaxel, fluorouracil
  - “Patient Population”: patients with cancer



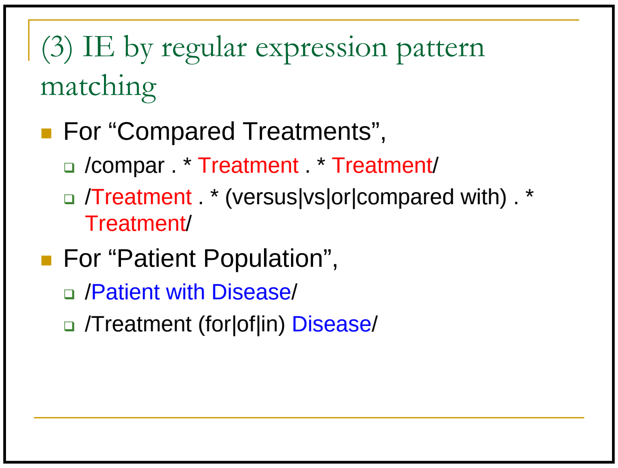
(1) Base NP chunking

- A base NP is defined as the shortest unit of noun phrase.
  - For example, “patients with cancer” is an NP but not a base NP.
- We use a SVM based chunker.
  - Training corpus is Penn Treebank.
  - Accuracy is around 90% in applying to our experiment.



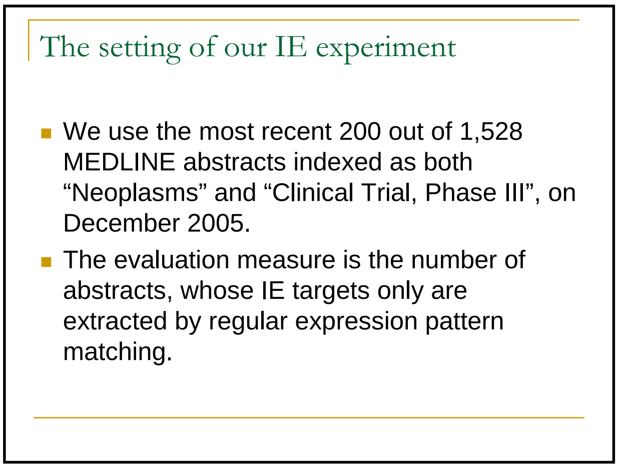
(2) Base NP categorization

- Attach a class label to each of base NPs.
  - We define class labels: “Disease”, “Treatment”, “Patient”, “Study”, “Others”.
- We use a SVM based categorizer.
  - Training corpus is our manually annotating clinical trial MEDLINE abstracts.
  - Accuracy is 70 ~ 90%.



(3) IE by regular expression pattern matching

- For “Compared Treatments”,
  - /compar . * Treatment . * Treatment/
  - /Treatment . * (versus|vs|or|compared with) . * Treatment/
- For “Patient Population”,
  - /Patient with Disease/
  - /Treatment (for|of|in) Disease/



The setting of our IE experiment

- We use the most recent 200 out of 1,528 MEDLINE abstracts indexed as both “Neoplasms” and “Clinical Trial, Phase III”, on December 2005.
- The evaluation measure is the number of abstracts, whose IE targets only are extracted by regular expression pattern matching.

Results of IE experiment

- For “Compared Treatments”,
  - we have successful results in 118 out of 200 abstracts.
- For “Patient Population”,
  - we have successful results in 125 out of 200 abstracts.
- Next, in order to improve the results, we conduct IE with filtering based on the document and sentence classification techniques.

IE with Document filtering and Sentence filtering

Document filtering: Motivations

- MEDLINE abstracts indexed as “Clinical Trial, Phase III” contain non phase III trials in fact.
  - For example, abstracts that just report the results of exploratory analyses using data or participants in past phase III trials are not excluded.
- The proportion of such kind of abstracts is about 30 %, that should be removed.

Document filtering: Methods and Results

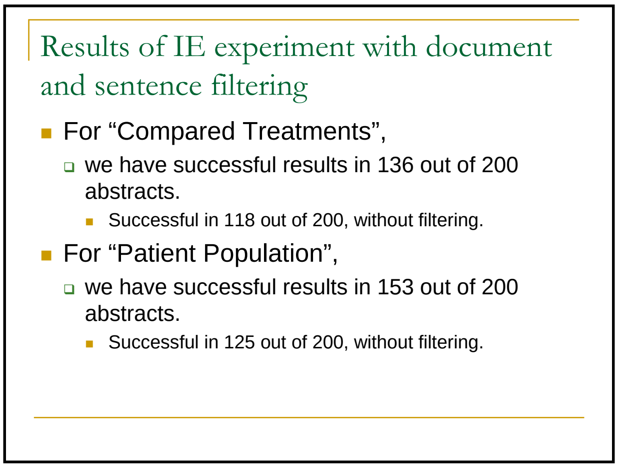
- Methods:
  - We use a SVM based document classifier.
  - Training corpus is our manually annotating clinical trial MEDLINE abstracts.
- Results:
  - Accuracy of filtering is around 90%.

Sentence filtering: Motivations

- Here, we are going to select the sentences that contain important keys.
- For example, a false positive sentence: “Subgroup analysis showed that patients with breast cancer had better survival”.
  - This is just the result of a subgroup analysis, which dose not provide firm evidence for EBM.

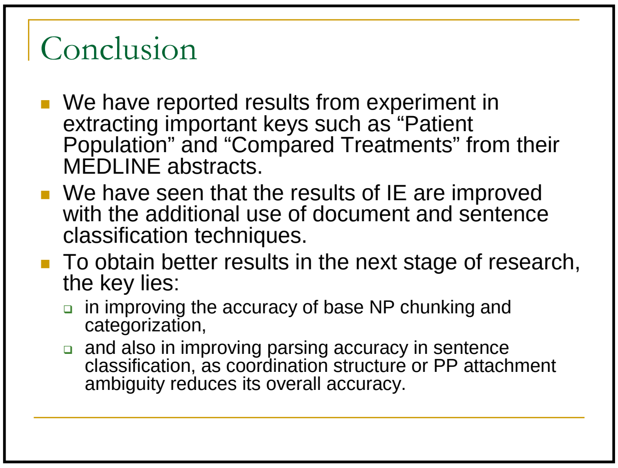
Sentence filtering: Methods and Results

- Methods:
  - We use a decision stumps and boosting based sentence classifier.
  - Training corpus is our manually annotating clinical trial MEDLINE abstracts.
- Results:
  - Accuracy of filtering is around 80%.



Results of IE experiment with document and sentence filtering

- For “Compared Treatments”,
  - we have successful results in 136 out of 200 abstracts.
    - Successful in 118 out of 200, without filtering.
- For “Patient Population”,
  - we have successful results in 153 out of 200 abstracts.
    - Successful in 125 out of 200, without filtering.



Conclusion

- We have reported results from experiment in extracting important keys such as “Patient Population” and “Compared Treatments” from their MEDLINE abstracts.
- We have seen that the results of IE are improved with the additional use of document and sentence classification techniques.
- To obtain better results in the next stage of research, the key lies:
  - in improving the accuracy of base NP chunking and categorization,
  - and also in improving parsing accuracy in sentence classification, as coordination structure or PP attachment ambiguity reduces its overall accuracy.