Ubiquitous Medicine - a trend in the medical community -
• This trend is supported by popularization of ubiquitous technology such as
  – Remote Diagnostic Imaging, or
  – 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 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

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".

A clinical trial result is always summarized in a table.

- A typical example (phase III)

<table>
<thead>
<tr>
<th>Endpoint (Efficacy)</th>
<th>Treatment A (New Drug)</th>
<th>Treatment B (Active Control)</th>
<th>statistical significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>value or score</td>
<td>value or score</td>
<td>p-value</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Endpoint (Safety)</th>
<th>frequency or count</th>
<th>frequency or count</th>
<th>p-value</th>
</tr>
</thead>
</table>

MEDLINE abstracts are just the rewriting of the result tables.

[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 versus 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.
- Information Extraction (IE) targets are:
  - “compared treatments”, “patient population”, “endpoints”, and so on.
- This can become a support for realizing EBM in the medical community.

Today’s presentation, we report ...

- Results of the two preliminary experiments for the summarization of clinical trial design information from MEDLINE abstracts.
  - Firstly, we used conventional Information Extraction (IE) methods to conduct an experiment in extraction of clinical trial design information.
  - Next, we performed sentence classification, using state-of-the-art sentence classification algorithm with the future goal of using those results to determine when to carry out IE.

Experiment I

Information Extraction (IE)

What is Information Extraction (IE) in general?

- The goal is to extract pre-specified types of events, entities or relationships from the documents.
- Extracted information is usually entered into a database,
  - for the purpose of analyzing the data for trends, giving a natural language summary, or simply serving for on-line access.
Again, our IE task:

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 targets:

1. Compared Treatment: peginterferon alfa-2a plus ribavirin versus interferon alfa-2a plus ribavirin
2. Endpoint: sustained virologic response
3. Patient Population: persons coinfected with HIV

We used conventional IE methods to estimate the difficulty of our task.

- Part-of-speech tagging
  - TnT tagger (Brants, 2000)
- Noun Phrase chunking
  - YamCha (Kudo and Matsumoto, 2001)
- Noun Phrase tagging
  - Manual labor using domain specific knowledge
- Extraction of IE targets by using manually written patterns

Part-of-speech tagging:

<table>
<thead>
<tr>
<th>TOKEN</th>
<th>Part-of-speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>We</td>
<td>PRP</td>
</tr>
<tr>
<td>conducted</td>
<td>VBD</td>
</tr>
<tr>
<td>a</td>
<td>DT</td>
</tr>
<tr>
<td>multi-center</td>
<td>NN</td>
</tr>
<tr>
<td>,</td>
<td></td>
</tr>
<tr>
<td>randomized</td>
<td>VBN</td>
</tr>
<tr>
<td>trial</td>
<td></td>
</tr>
<tr>
<td>comparing</td>
<td>VBG</td>
</tr>
</tbody>
</table>

Noun Phrase chunking:

<table>
<thead>
<tr>
<th>TOKEN</th>
<th>POS</th>
<th>NP chunk</th>
</tr>
</thead>
<tbody>
<tr>
<td>We</td>
<td>PRP</td>
<td>B (Begin)</td>
</tr>
<tr>
<td>conducted</td>
<td>VBD</td>
<td>O (Outside)</td>
</tr>
<tr>
<td>a</td>
<td>DT</td>
<td>B (Begin)</td>
</tr>
<tr>
<td>multi-center</td>
<td>NN</td>
<td>I (Inside)</td>
</tr>
<tr>
<td>,</td>
<td></td>
<td>I (Inside)</td>
</tr>
<tr>
<td>randomized</td>
<td>VBN</td>
<td>I (Inside)</td>
</tr>
<tr>
<td>trial</td>
<td>NN</td>
<td>I (Inside)</td>
</tr>
<tr>
<td>comparing</td>
<td>VBG</td>
<td>O (Outside)</td>
</tr>
</tbody>
</table>

Noun Phrase tagging:

- Manual labor using domain specific knowledge

NP tagged sentence by YamCha:

“[We] conducted [a multi-center, randomized trial] comparing [peginterferon plus ribavirin] with [interferon plus ribavirin] for [the treatment of] [chronic hepatitis C].”

NP tagging:

“[We] conducted [STUDY] comparing [DRUG] with [DRUG] for [THERAPY] of [DISEASE].”

Our Noun Phrase tag set:

- **DISEASE:** disease, symptom, virus
- **DRUG:** drug, chemical compound
- **STUDY:** clinical trial
- **THERAPY:** treatment, regimen
- **PATIENT:** participants in the trial
- **TARGET:** endpoints
- **SCHEDULE:** time schedule of the trial
- **VALUE:** value of TARGET
- **NUMBER:** numeral expression

- **Example:**
  - chronic hepatitis C
  - interferon
  - clinical trial
  - antiviral treatment
  - HBeAg-positive patients
  - efficacy and safety
  - an additional 24 weeks
  - significantly higher rates
  - 20 percent
Extraction of IE targets:
- using our manually written patterns -

A regular expression pattern for “Compared Treatment”:
“comparing [DRUG] with [DRUG]”

“[We] conducted [STUDY] comparing [DRUG] with [DRUG] ...”

IE Result:
- Compared Treatment: “peginterferon plus ribavirin”
- Compared Treatment: “interferon plus ribavirin”

Data used in our experiment:
- We downloaded the 50 most recent abstracts of clinical trials from the MEDLINE database on October 2004.
- To simplify the experiment, abstracts were selected from the medical area of hepatitis.

Evaluation Metrics in our IE experiment:
- We used four measures.
- For IE target entities,
  - Precision for entity extraction (Ent_pre)
  - Recall for entity extraction (Ent_rec)
- For Abstracts,
  - Precision for abstract summarization (Sum_pre)
  - Recall for abstract summarization (Sum_rec)

Results:
- We show two types of results:
  - IE from titles alone, and
  - IE from titles and main texts.
- The results from titles alone can be considered as the baseline,
  - because just putting together the titles is close to summarizing the articles.

Results of IE from titles alone and from titles and main texts:

<table>
<thead>
<tr>
<th></th>
<th>Compared Treatment</th>
<th>Endpoint</th>
<th>Patient Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ent_pre</td>
<td>76.9%</td>
<td>66.7%</td>
<td>88.2%</td>
</tr>
<tr>
<td>Ent_rec</td>
<td>50.2%</td>
<td>29.0%</td>
<td>53.6%</td>
</tr>
<tr>
<td>Sum_pre</td>
<td>86.0%</td>
<td>96.0%</td>
<td>94.0%</td>
</tr>
<tr>
<td>Sum_rec</td>
<td>40.0%</td>
<td>24.0%</td>
<td>50.0%</td>
</tr>
<tr>
<td></td>
<td>Ent_pre</td>
<td>71.4%</td>
<td>71.9%</td>
</tr>
<tr>
<td></td>
<td>Ent_rec</td>
<td>78.3%</td>
<td>59.4%</td>
</tr>
<tr>
<td></td>
<td>Sum_pre</td>
<td>70.0%</td>
<td>82.0%</td>
</tr>
<tr>
<td></td>
<td>Sum_rec</td>
<td>66.0%</td>
<td>52.0%</td>
</tr>
</tbody>
</table>

Why performance isn’t good?
- The patterns based on heuristics have no theoretical guarantee that they are correct.
- In the next, we show experimental results of sentence classification that might overcome the difficulties found in this IE experiment.
Experiment II
Sentence Classification

Our Sentence Classification task:
- Classifying sentences into “yes” or “no” with respect to whether the sentence includes IE targets or not.

NP tagged sentence:

Classification Result:
- Compared Treatment: +1 (Yes)
- Endpoint: -1 (No)
- Patient Population: +1 (Yes)

Why Sentence Classification?
- We hypothesize that the IE accuracy can be increased
  - by classifying sentences and only performing IE on sentences that are most likely to contain relevant information.
  - Furthermore, classifying or filtering sentences could be the rational step to get to our goal
    - in the respect that our final goal is summarization of clinical trial design information.

We used BACT, a state-of-the-art sentence classification algorithm.
- Up to Noun Phrase tagging, the experiment process is same as IE experiment.
- We used BACT (Kudo and Matsumoto, 2004)
  - to acquire optimal classification patterns by machine learning, and
  - to classify sentences according to those automatically constructed patterns.

BACT learns from training data as ordered trees.
- In case of Bag-of-words (BOW) assumption:

  An example sentence:
  “[We] conducted [STUDY] comparing …”

BACT learns from training data as ordered trees.
- In case of N-gram assumption:

  An example sentence:
  “[We] conducted [STUDY] comparing …”
BACT learns from training data as ordered trees.
• In case of dependency grammar restriction:

An example sentence:
“[We] conducted [STUDY] comparing [DRUG] with [DRUG] …”

BACT searches for sub-trees in each ordered tree.
• Searches all sub-trees comprehensively.

BACT ranks the sub-trees (calculates weights):
• An example of dependency grammar restriction.

<table>
<thead>
<tr>
<th>automatically constructed pattern by BACT that include “DRUG”</th>
<th>Compared Treatment</th>
<th>Endpoint</th>
<th>Patient Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>“PATIENT received DRUG”</td>
<td>0.048</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>“DRUG”</td>
<td>0.046</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>“TARGET of DRUG”</td>
<td>–</td>
<td>0.035</td>
<td>–</td>
</tr>
<tr>
<td>“DRUG, DRUG”</td>
<td>0.013</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>“received DRUG”</td>
<td>0.01</td>
<td>0.023</td>
<td>–</td>
</tr>
<tr>
<td>“of DRUG”</td>
<td>0.006</td>
<td>0.012</td>
<td>–</td>
</tr>
<tr>
<td>“with DRUG”</td>
<td>-0.004</td>
<td>–</td>
<td>-0.026</td>
</tr>
<tr>
<td>“to DRUG”</td>
<td>-0.013</td>
<td>–</td>
<td>-0.012</td>
</tr>
<tr>
<td>“in DRUG”</td>
<td>-0.015</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

BACT classifies sentences according to automatically constructed patterns.

NP tagged sentence:

Classification Result:
Compared Treatment: +1 Yes
Endpoint: -1 No
Patient Population: +1 Yes

Evaluation Metrics:
• We used two measures.
  • Precision = tp / (tp + fp)
  • Recall = tp / (tp + tn)
  – tp means true positive, fp means false positive, and tn means true negative.
  – Precision is the correctness of the system when it classifies sentences to “yes”.
  – Recall is the proportion of “yes” sentences that the system classifies to “yes”.

Data used in our experiment:
• Same data as IE experiment.
• We downloaded the 50 most recent abstracts of clinical trials from the MEDLINE database on October 2004.
• To simplify the experiment, abstracts were selected from the medical area of hepatitis.
Sentence Classification experiment

Results (five-fold cross validation):

<table>
<thead>
<tr>
<th></th>
<th>Compared Treatment</th>
<th>Endpoint</th>
<th>Patient Population</th>
</tr>
</thead>
<tbody>
<tr>
<td># total sentence</td>
<td>562</td>
<td>562</td>
<td>562</td>
</tr>
<tr>
<td># total “yes” sentence</td>
<td>90</td>
<td>76</td>
<td>55</td>
</tr>
<tr>
<td>BOW</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>precision</td>
<td>82.3%</td>
<td>71.7%</td>
<td>81.8%</td>
</tr>
<tr>
<td>recall</td>
<td>70.8%</td>
<td>64.7%</td>
<td>69.3%</td>
</tr>
<tr>
<td>N-gram</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>precision</td>
<td>82.6%</td>
<td>81.5%</td>
<td>81.5%</td>
</tr>
<tr>
<td>recall</td>
<td>71.7%</td>
<td>64.7%</td>
<td>81.5%</td>
</tr>
<tr>
<td>dependency</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>precision</td>
<td>86.8%</td>
<td>75.2%</td>
<td>81.5%</td>
</tr>
<tr>
<td>recall</td>
<td>78.5%</td>
<td>71.4%</td>
<td>71.4%</td>
</tr>
</tbody>
</table>

Today’s presentation, we have reported …

• Results of the two preliminary experiments to estimate the difficulty of our task.
• These preliminary experiments show that the combination of IE methodology and sentence classification can be the solution to the summarization task in clinical trial MEDLINE abstracts.
• So we plan to construct a complete pipeline from sentence classification to IE.

Future Work

• Construction of bigger corpora.
• Automate NP tagging.
• In the IE subtask,
  – identification of correspondence between entities and mentions.
• In the subtask of sentence classification using BACT,
  – improving parsing accuracy such that come from coordination structure or PP attachment ambiguity.

Thank you!