Text Mining for Knowledge Discovery and Ontology Extension

Jong C. Park
KAIST, Korea
park@nlp.kaist.ac.kr
nlp.kaist.ac.kr/~park

Workshop on Text Mining, Ontology and Natural Language Processing in Biomedicine
March 20-21, 2006
Overview

1. Introduction

2. Applications for Biologists
   - Information Extraction
   - Knowledge Induction
   - Convenient Interface

3. Conclusion
Introduction: Techniques for Text Mining in Biomedicine

Many techniques of NLP are utilized.

Grammar-based parsing
- CFG, HPSG, CCG …

Statistical approach
- Co-occurrence …

Pattern matching
- Regular expression

Machine learning
- HMM, SVM, MEM …
Text mining is helpful in many areas.

Text Mining for Knowledge Discovery and Ontology Extension

March 21, 2006
Introduction: Towards Better Performance of Text Mining

Text mining may require sophisticated analyses.

Term variations in the literature

- **Word level**
  - Morphological variations
  - Semantic variations: Synonyms, Hyponyms

- **Syntax level**
  - The left-to-right order of component words may not always be the same as that of a corresponding term.

- **Discourse level**
  - Component words of a term may be distributed across multiple sentences.

March 21, 2006

Text Mining for Knowledge Discovery and Ontology Extension
Research Directions of Our Group at KAIST

NLP techniques for Bioinformatics

Convenient Interface

BiopathwayBuilder
Visualization of molecular interaction

BioNLQ
NL query for heterogeneous DB

Summarization
Generation of gene summary

Knowledge Induction

AutoGO
Automatic extension of Gene Ontology

BioContrasts
Automatic extraction of contrastive information

BioIE
Automatic extraction of p-p interaction

Information Extraction
Information Extraction

Bio I E

BioContrasts

Jung-jae Kim
jjkim@nlp.kaist.ac.kr
nlp.kaist.ac.kr/~jjkim
Information extraction of biological relation

- Extraction of general biological interactions of arbitrary types, including protein-protein interaction

Characteristics of BioIE

- Unknown word handling
- Semantic class identification
- Analysis of important linguistic constructions such as acronyms, appositive structures, and anaphoric expressions
Biol E: Example Procedure

MEDLINE abstract

Keyword detection + Pattern matching (A inhibited B)

... inhibited ...

Selection of multiple pairs of candidate arguments (A, B)

As assessed by [a genetic assay that measures [AAI-dependent DNA binding, [TraM]]] inhibited [[[TraR function] before and after the transcription factor] had bound to its DNA recognition site].

Parsing for grammaticality inspection with CCG

As assessed by a genetic assay that measures AAI-dependent DNA binding, [TraM] inhibited [TraR function] before and after the transcription factor had bound to its DNA recognition site.
### Biol E: Experimental Results

<table>
<thead>
<tr>
<th>Keyword Set</th>
<th>Test Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>K1</strong></td>
<td></td>
</tr>
<tr>
<td>interact</td>
<td><strong>topoisomerase</strong></td>
</tr>
<tr>
<td>bind</td>
<td>inhibitors</td>
</tr>
<tr>
<td>associate</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td><strong>K2</strong></td>
<td></td>
</tr>
<tr>
<td>activate</td>
<td></td>
</tr>
<tr>
<td>inhibit</td>
<td></td>
</tr>
<tr>
<td>stimulate</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>K1</th>
<th>K2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>92.1</td>
<td>56.9</td>
</tr>
<tr>
<td>Recall</td>
<td>70.3</td>
<td>70.3</td>
</tr>
<tr>
<td>F-score</td>
<td>67.7</td>
<td>67.7</td>
</tr>
</tbody>
</table>

---

March 21, 2006

Text Mining for Knowledge Discovery and Ontology Extension
Contrasts

- Richly informative units of linguistic expressions from the biomedical literature
- Express explicit difference and implicit similarity in functions

BioContrasts

- Existing protein-protein interaction databases capture mostly positive and individual relations
- Current text mining work has also focused on extracting positive and individual relationships
- Our database enables biologists to exploit such a rich resource of contrastive information already available in the literature
BioContrasts: Example Procedure

“both eIF-4B and eIF-4F, but not eIF-4A, interact with ribosomes in the presence of specific factors and ATP”

Identify Coordination

Expression for positive object: “both eIF-4B and eIF-4F”
Expression for negative object: “eIF-4A”

Extract Contrastive Protein Names

Positive protein names: “eIF-4B”, “eIF-4F”
Negative protein names: “eIF-4A”

Ground Protein Names

Positive Swiss-Prot entries: “IF4B_HUMAN”, “IF4B_YEAST”, “IF4F1_YEAST”, “IF4F2_YEAST”
Negative Swiss-Prot entries: “IF4A_SCHPO”, “IF4A_YEAST”

Extract Presupposed Property

“X interact with ribosomes”, where
X = “IF4B_YEAST” or “IF4F1_YEAST” or “IF4F2_YEAST” and X ≠ “IF4A_YEAST”
BioContrasts: Results

- System implemented in Python
  - Enhanced performance: 0.038 sec/abstract

- Extraction corpus
  - 2.5 million ‘not’-containing MEDLINE abstracts

- Contrast extraction
  - 799,169 pairs of contrastive expressions
  - 11,284 pairs of contrastive protein names
  - 41,471 contrast btw Swiss-Prot entries

- A web-portal for public access

March 21, 2006
BioContrasts: Evaluation

Evaluation set

- 100 randomly selected protein-protein contrasts

Indicative results

- 97/100 (97.0% precision) on extraction of contrastive protein names
- 164/182 (90.1% precision) on protein name grounding (18 names without clarifying phrases)
Huntington’s disease

Propose non-obvious (e.g. non-homologous) candidate proteins in incomplete pathways
The pathway of Huntington’s disease (HD) involves the neurotrophic factor (BDNF).

Find any contrast involving a protein member

BioContrasts includes a contrast between BDNF and CNTF, another neurotrophic factor, with the evidence of PMID:8584263.

MEDLINE search

A MEDLINE search with “BDNF CNTF Huntington’s disease” leads us to find an abstract (PMID:12062094).

Identify relation of the candidate with the pathway

For example, we have shown using an in vitro neuronal model of HD that CNTF and BDNF block polyQ-huntingtin-induced cell death. In vivo, CNTF has also been shown to be neuro-protective in rats and monkeys following excitotoxic lesions that reproduce HD. (PMID:12062094)
BioContrasts: Web-Portal for Public Access

http://biocontrasts.biopathway.org
http://biocontrasts.i2r.a-star.edu.sg

KEGG pathway: hsa05040 (Huntington's disease)

Application 1-1. Identify contrastive pathway members that may have different roles

<table>
<thead>
<tr>
<th>Original KEGG Member</th>
<th>Contrastive Member 1</th>
<th>Contrastive Member 2</th>
<th>Evidence</th>
<th>Relevant Literature Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASP</td>
<td>CASP3_HUMAN (Caspase-3 precursor)</td>
<td>CASP1_HUMAN (Caspase-1 precursor)</td>
<td>1</td>
<td>PUBMED</td>
</tr>
<tr>
<td>CASP</td>
<td>CASP3_HUMAN (Caspase-3 precursor)</td>
<td>CASP1_HUMAN (Caspase-1 precursor)</td>
<td>1</td>
<td>PUBMED</td>
</tr>
<tr>
<td>Clathrin</td>
<td>CLCA_HUMAN (Clathrin light chain A)</td>
<td>CLCB_HUMAN (Clathrin light chain B)</td>
<td>1</td>
<td>PUBMED</td>
</tr>
</tbody>
</table>

Application 1-2. Suggest candidate members of KEGG pathway

<table>
<thead>
<tr>
<th>Candidate Member</th>
<th>Related KEGG Member</th>
<th>Relevant Literature Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT5_HUMAN (Neurotrophin-5 precursor)</td>
<td>BDNF_HUMAN (Brain-derived neurotrophic factor precursor)</td>
<td>PUBMED</td>
</tr>
<tr>
<td>CNTF_HUMAN (Ciliary neurotrophic factor)</td>
<td>BDNF_HUMAN (Brain-derived neurotrophic factor precursor)</td>
<td>PUBMED</td>
</tr>
<tr>
<td>FGF2_HUMAN (Heparin-binding growth factor 2 precursor)</td>
<td>BDNF_HUMAN (Brain-derived neurotrophic factor precursor)</td>
<td>PUBMED</td>
</tr>
<tr>
<td>CASP_HUMAN</td>
<td>BDNF_HUMAN</td>
<td>PUBMED</td>
</tr>
</tbody>
</table>

March 21, 2006
Knowledge Induction

AutoGO

Summarization

Jin-Bok Lee
jblee@nlp.kaist.ac.kr
nlp.kaist.ac.kr/~jblee
Extension of Gene Ontology

- Prediction of more detailed terms of GO
- Induction of the terms with context-sensitive rules from syntactic relations among the existing terms

Validation of the extended GO

- Validation of the candidate terms with identification of the terms
- Flexible identification of terms using syntactic dependencies in the literature
AutoGO: Extension of GO

Comparison among 3 versions of GO

- Automatically predicted concepts that are newly introduced by domain experts in the more recent version of GO.
- Automatically predicted concepts that are NOT introduced by domain experts yet.
- Modification by domain experts
  - Discarded concepts
- Not predicted by our system
To identify molecules regulating this interaction, we generated FDC-staining monoclonal antibodies (mAbs) and screened them for their ability to block FDC-mediated costimulation of growth and differentiation of CD40-stimulated B cells.

(PMID:10727470)
# AutoGO: Evaluation of Validation

## Experimental results of candidate term validation

<table>
<thead>
<tr>
<th>Confirmed</th>
<th>Sentence validation</th>
<th>Abstract validation</th>
<th>Verified terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>71.9 %</td>
<td>62.7 %</td>
<td>58.2 %</td>
</tr>
<tr>
<td>Recall</td>
<td>82.1 %</td>
<td>61.0 %</td>
<td>(69.1 %)</td>
</tr>
<tr>
<td>Not confirmed</td>
<td>Precision</td>
<td>52.4 %</td>
<td>51.7 %</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>84.6 %</td>
<td>(50.0 %)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
AutoGO: Results of GO Extension

Sample part of the extended GO

- Imaginal disc morphogenesis (549)
  - Clypeo-labral disc morphogenesis (0)
  - Eye-antennal disc morphogenesis (353)
  - Genital disc morphogenesis (13)
  - Haltere disc morphogenesis (1)
  - Imaginal disc eversion (4)
  - Imaginal disc fusion (9)
  - Labial disc morphogenesis (0)
  - Leg disc morphogenesis (39)
  - Morphogenesis of larval imaginal disc epithelium (22)
  - Prothoracic disc morphogenesis (0)
  - Regulation of imaginal disc morphogenesis
  - Wing disc morphogenesis (192)
**Summarization**

- **Gene summary**
  - An effective way to grasp new biological concepts

- **Informative summary**
  - Concept ranking and contrastive information acquisition

- **Coherent summary**
  - Discourse planning and sentence linking
Convenient Interface

BioNLQ

BiopathwayBuilder
BioNLQ

Data search with a natural language query

- Conceptual data search
- Multiple search paths for relevant data
- Unified access to multiple databases

Explorative search for biological data

- Diverse information access with ranking
- Accurate results and broad range of results
- Guidance over data retrieval process
Roles of a visualization system

- To complement the functions of an IE system by guiding the user in his/her inference process over the extracted facts.
- To understand data of a high complexity and to lead to consequent knowledge discovery.

Customized visualization

- Customized view using semantic classification
- Layered visualization for representing different levels of detail
BiopathwayBuilder

query → MEDLINE → raw documents → information extraction → filtering → molecular interaction database → distortion → layered visualization for representing different levels of detail

level of detail

user

data manipulation

data analysis

March 21, 2006

Text Mining for Knowledge Discovery and Ontology Extension
BiopathwayBuilder: Visualization Tool

Intuitive interface for molecular interactions

Yeast data set (235 abstracts)

Keywords: protein, gene, drug, interaction, S. cerevisiae
BiopathwayBuilder: Toward Knowledge Discovery

Customized view

All the layers are shown. 

Lower layer is hidden.
Summary

Text Mining
- Techniques for Text Mining in Biomedicine
- Applications of Text Mining
- Towards Better Performance of Text Mining

Applications for Biologists
- Information Extraction
- Knowledge Induction
- Convenient Interface
Thank You!