Information Extraction from Scientific Texts

Junichi Tsujii
Graduate School of Science
University of Tokyo
Japan
Texts are one of the major sources of information and knowledge.

However, they are not transparent.

They have to be systematically integrated with the other sources like data bases, numerical data, etc.

Natural Language Processing--IE
Overview of GENIA System

Corpus Module
- Markup generation / compilation
- Annotated corpus construction

Information Extraction Module
- Identify & classify terms
- Identify events

Retrieval Module
- Request enhancement
- Spawn request
- Classify documents

Interface Module
- GUI
- HTML conversion
- System integration

Database Module
- DB design / access / management
- DB construction

User
- IR Request
- Abstract
- Full Paper

Background Knowledge
- Ontology
- Markup language
- Data model

Concept Module
- BK design / construction / compilation

Corpus
- Raw(OCR)
- Text Structure
- Annotated

Database
- Document
- Named-Entity
- Event

MEDLINE

Security
Plan

1. What is IE?

2. General Framework of NLP

3. Basic IE techniques

4. IE in Biology

Automatic Term Recognition
(S. Ananiadou)
What is IE?
Application Tasks of NLP

(1) Information Retrieval/Detection
    To search and retrieve documents in response to queries for information

(2) Passage Retrieval
    To search and retrieve part of documents in response to queries for information

(3) Information Extraction
    To extract information that fits pre-defined database schemas or templates, specifying the output formats

(4) Question/Answering Tasks
    To answer general questions by using texts as knowledge base: Fact retrieval, combination of IR and IE

(5) Text Understanding
    To understand texts as people do: Artificial Intelligence
Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 iron and “metal wood” clubs a month.

Example of IE: FASTUS(1993)

TIE-UP-1
Relationship: TIE-UP
Entities: “Bridgestone Sport Co.”
“a local concern”
“a Japanese trading house”
Joint Venture Company:
“Bridgestone Sports Taiwan Co."
Activity: ACTIVITY-1
Amount: NT$200000000

ACTIVITY-1
Activity: PRODUCTION
Company:
“Bridgestone Sports Taiwan Co.”
Product:
“iron and ‘metal wood’ clubs”
Start Date:
DURING: January 1990
FASTUS

Based on finite states automata (FSA)

1. Complex Words:
   Recognition of multi-words and proper names

2. Basic Phrases:
   Simple noun groups, verb groups and particles

3. Complex phrases:
   Complex noun groups and verb groups

4. Domain Events:
   Patterns for events of interest to the application
   Basic templates are to be built.

5. Merging Structures:
   Templates from different parts of the texts are merged if they provide information about the same entity or event.
Interpretation of Texts

(1) Information Retrieval/Detection

User

(2) Passage Retrieval

User

(3) Information Extraction

System

(4) Question/Answering Tasks

System

(5) Text Understanding

System
<table>
<thead>
<tr>
<th>Task</th>
<th>Clarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Retrieval/Detection</td>
<td>Rather clear</td>
</tr>
<tr>
<td>Passage Retrieval</td>
<td>A bit vague</td>
</tr>
<tr>
<td>Information Extraction</td>
<td>Rather clear</td>
</tr>
<tr>
<td>Question/Answering Tasks</td>
<td>A bit vague</td>
</tr>
<tr>
<td>Text Understanding</td>
<td>Very vague</td>
</tr>
</tbody>
</table>
Query

Collection of Documents

Precision: $\frac{C}{M}$

Recall: $\frac{C}{N}$

F-Value: $\frac{2P \cdot R}{P + R}$

N: Correct Templates
M: Retrieved Templates
C: Correct Templates that are actually retrieved

More complicated due to partially filled templates
General Framework of NLP
General Framework of NLP

John runs.

John run+s.

P-N  V  3-pre
N  plu

Pred: RUN
Agent: John

John is a student.
He runs.
General Framework of NLP

Tokenization

Morphological and Lexical Processing

Part of Speech Tagging

Syntactic Analysis

Inflection/Derivation

Semantic Analysis

Compounding

Context processing Interpretation

Term recognition

Domain Analysis

Appelt:1999
Difficulties of NLP

(1) Robustness: General Framework of NLP
Incomplete Knowledge

- Morphological and Lexical Processing
- Syntactic Analysis
- Semantic Analysis
- Context processing
- Interpretation

Incomplete Lexicons
- Open class words
- Terms
- Term recognition
- Named Entities
- Company names
- Locations
- Numerical expressions
Difficulties of NLP

(1) Robustness: General Framework of NLP
Incomplete Knowledge

Incomplete Grammar
Syntactic Coverage
Domain Specific Constructions
Ungrammatical Constructions

Morphological and Lexical Processing

Syntactic Analysis

Semantic Analysis

Context processing Interpretation
Difficulties of NLP

(1) Robustness: General Framework of NLP
Incomplete Knowledge

Predefined Aspects of Information

- Morphological and Lexical Processing
- Syntactic Analysis
- Semantic Analysis
- Context processing
- Interpretation

Incomplete Domain Knowledge Interpretation Rules
Difficulties of NLP

(1) Robustness: Incomplete Knowledge

(2) Ambiguities: Combinatorial Explosion

Most words in English are ambiguous in terms of their part of speeches.

- runs: v/3pre, n/plu
- clubs: v/3pre, n/plu

and two meanings
Difficulties of NLP

(1) Robustness: General Framework of NLP
Incomplete Knowledge

(2) Ambiguities:
- Combinatorial Explosion
- Structural Ambiguities
- Predicate-argument Ambiguities

Diagram:
- Morphological and Lexical Processing
- Syntactic Analysis
- Semantic Analysis
- Context processing Interpretation
Structural Ambiguities

(1) Attachment Ambiguities

John bought a car with large seats.
John bought a car with $3000.

The manager of Yaxing Benz, a Sino-German joint venture
The manager of Yaxing Benz, Mr. John Smith

(2) Scope Ambiguities

young women and men in the room

(3) Analytical Ambiguities

Visiting relatives can be boring.

Semantic Ambiguities (1)

John bought a car with Mary.
$3000 can buy a nice car.

Semantic Ambiguities (2)

Every man loves a woman.

Co-reference Ambiguities
Difficulties of NLP

(1) Robustness: Incomplete Knowledge

(2) Ambiguities:
- Combinatorial Explosion
- Structural Ambiguities
- Predicate-argument Ambiguities

General Framework of NLP
- Morphological and Lexical Processing
- Syntactic Analysis
- Semantic Analysis
- Context processing
- Interpretation
Note:

Ambiguities vs Robustness

More comprehensive knowledge: More Robust
big dictionaries
comprehensive grammar

More comprehensive knowledge: More ambiguities

Adaptability: Tuning, Learning
Framework of IE

IE as compromise NLP
Techniques in IE

(1) **Domain Specific Partial Knowledge:**
Knowledge relevant to information to be extracted

(2) **Ambiguities:**
Ignoring irrelevant ambiguities
Simpler NLP techniques

(3) **Robustness:**
Coping with Incomplete dictionaries
(open class words)
Ignoring irrelevant parts of sentences

(4) **Adaptation Techniques:**
Machine Learning, Trainable systems
General Framework of NLP

Morphological and Lexical Processing

Syntactic Analysis

Semantic Analysis

Context Processing

Interpretation

Open class words:
- Named entity recognition
  - (ex) Locations
    - Persons
    - Companies
    - Organizations
    - Position names

Domain specific rules:
- &lt;Word&gt;&lt;Word&gt;, Inc.
- Mr. &lt;Cpt-L&gt;. &lt;Word&gt;

Machine Learning:
- HMM, Decision Trees
- Rules + Machine Learning

Local Context

Statistical Bias

95 %

FSA rules

Statistic taggers

F-Value

90

Domain Dependent
General Framework of NLP

- Morphological and Lexical Processing
- Syntactic Analysis
- Semantic Analysis
- Context processing
- Interpretation

**FASTUS**

Based on finite states automata (FSA)

1. Complex Words:
   Recognition of multi-words and proper names

2. Basic Phrases:
   Simple noun groups, verb groups and particles

3. Complex phrases:
   Complex noun groups and verb groups

4. Domain Events:
   Patterns for events of interest to the application
   Basic templates are to be built.

5. Merging Structures:
   Templates from different parts of the texts are merged if they provide information about the same entity or event.
Chomsky Hierarchy of Grammar

Regular Grammar
Context Free Grammar
Context Sensitive Grammar
Type 0 Grammar

Hierarchy of Automata

Finite State Automata
Push Down Automata
Linear Bounded Automata
Turing Machine

Computationally more complex, Less Efficiency
John’s interesting book with a nice cover
Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 “metal wood” clubs a month.

**Example of IE: FASTUS(1993)**

1. Complex words
   - a Japanese tea house
   - a [Japanese tea] house
   - a Japanese [tea house]

2. Basic Phrases:
   - Bridgestone Sports Co.: Company name
   - said: Verb Group
   - Friday: Noun Group
   - it: Noun Group
   - had set up: Verb Group
   - a joint venture: Noun Group
   - in: Preposition
   - Taiwan: Location
Bridgestone Sports Co. said Friday it had set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be supplied to Japan.

The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990 with production of 20,000 “metal wood” clubs a month.
Example of IE: FASTUS(1993)

[COMPANY] said Friday it [SET-UP] [JOINT-VENTURE] in [LOCATION] with [COMPANY] to produce [PRODUCT] to be supplied to [LOCATION].


2. Basic Phrases:
   Bridgestone Sports Co.: Company name
   said : Verb Group
   Friday : Noun Group
   it : Noun Group
   had set up : Verb Group
   a joint venture : Noun Group
   in : Preposition
   Taiwan : Location

3. Complex Phrases
   Syntactic structures relevant to information to be extracted are dealt with.
Syntactic variations

GM set up a joint venture with Toyota.
GM announced it was setting up a joint venture with Toyota.
GM signed an agreement setting up a joint venture with Toyota.
GM announced it was signing an agreement to set up a joint venture with Toyota.

GM plans to set up a joint venture with Toyota.
GM expects to set up a joint venture with Toyota.
Example of IE: FASTUS(1993)

[COMPANY] [SET-UP] [JOINT-VENTURE] in [LOCATION] with [COMPANY] to produce [PRODUCT] to be supplied to [LOCATION].


3. Complex Phrases
4. Domain Events

[COMPANY][SET-UP][JOINT-VENTURE] with [COMPANY]
[COMPANY][SET-UP][JOINT-VENTURE] (others)* with [COMPANY]

The attachment positions of PP are determined at this stage. Irrelevant parts of sentences are ignored.
Complications caused by syntactic variations

Relative clause

The mayor, who was kidnapped yesterday, was found dead today.

[NG] Relpro {NG/others}* [VG] {NG/others}*[VG]
[NG] Relpro {NG/others}* [VG]

Basic patterns → Surface Pattern Generator → Patterns used by Domain Event

Relative clause construction
Passivization, etc.
FASTUS

Based on finite states automata (FSA)

1. Complex Words:  
NP, who was kidnapped, was found.

2. Basic Phrases:

3. Complex phrases:

4. Domain Events:  
Patterns for events of interest to the application  
Basic templates are to be built.

5. Merging Structures:
Templates from different parts of the texts are merged if they provide information about the same entity or event.

Piece-wise recognition of basic templates

Reconstructing information carried via syntactic structures by merging basic templates
IE in Biology
CSNDB
(National Institute of Health Sciences)

• A data- and knowledge- base for signaling pathways of human cells.
  – It compiles the information on biological molecules, sequences, structures, functions, and biological reactions which transfer the cellular signals.
  – Signaling pathways are compiled as binary relationships of biomolecules and represented by graphs drawn automatically.
  – CSNDB is constructed on ACEDB and inference engine CLIPS, and has a linkage to TRANSFAC.
  – Final goal is to make a computerized model for various biological phenomena.
Example. 1

• A Standard Reaction

Signal_Reaction:

“EGF receptor $\rightarrow$ Grb2”

From_molecule “EGF receptor”
To_molecule “Grb2”
Tissue “liver”
Effect “activation”
Interaction
“SH2+phosphorylated Tyr”
Reference [Yamauchi_1997]
Example. 3

- A Polymerization Reaction

Signal_Reaction:

"Ah receptor + HSP90 \rightarrow "

- Component "Ah receptor" "HSP90"
- Effect "activation dissociation"
- Interaction
  "PAS domain"
  "of Ah receptor"
- Activity
  "inactivation of Ah receptor"
- Reference [Powell-Coffman_1998]
An active phorbol ester must therefore, presumably by activation of protein kinase C, cause dissociation of a cytoplasmic complex of NF-kappa B and I kappa B by modifying I kappa B.

**E1**: An active phorbol ester activates protein kinase C.

**E2**: The active phorbol ester modifies I kappa B.

**E3**: It dissociates a cytoplasmic complex of NF-kappa B and I kappa B.
Full parser based on good grammar formalisms

1. Several attempts of using full parsers:
   To improve the Precision

2. Systematic treatment of interaction of the different phases:
   Unification-based grammar formalisms

The two papers in the NLP session of PSB 2001
Experiment
(A.Yakushiji et.al, PSB2001)

XHPSG: HPSG-like Grammar translated from
XTAG of U-Penn (Y.Tateishi, TAG+ workshop 98)
Automatic conversion: Detailed, empirical comparison of
grammars of different formalisms (+LFG)
Terms (Compound nouns) are chunked beforehand.

180 sentences from abstracts in MEDLINE

The average parse time per sentence: 2.7 sec by a naïve parser
(This can be improved by the multi-stage parser by 50 times)
Argument Frame Extractor

133 argument structures, marked by a domain specialist in 97 sentences among the 180 sentences

<table>
<thead>
<tr>
<th>Parsing Failures</th>
<th>Extractable from pp’s</th>
<th>Not extractable</th>
<th>Memory limitation, etc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extracted Uniquely</td>
<td>31</td>
<td>26</td>
<td>68%</td>
</tr>
<tr>
<td>Extracted with ambiguity</td>
<td>32</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>Not extractable</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory limitation, etc</td>
<td>17</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ontology: Knowledge of the Domain

Open class words:
   Named entity recognition
   (ex) Locations
   Persons
   Companies
   Organizations
   Position names

More refined semantic classes with part-whole relationships, properties, Etc.

Acronyms, variants, Etc.
Bio Term Bank

- A database for all sort of biological terms collected from genome databases and biological texts.
- It will contain 2 million terms in 2001 and 5 million terms until 2005.
- Terms are classified by biochemical and terminological attributes, grounded on their resources.

Biological ontology committee Japan organized by T. Takagi and T. Takai, U.Tokyo in Genome Projects of MESSC (2000.4 ~ 2005.3)
GENIA ontology
(current version)

+-name+-source+-natural+-organism+-multi-cell organism
  |        |         |          +-mono-cell organism
  |        |         |          +-virus
  |        |         |          +-tissue
  |        |         |          +-cell type
  |        |         |          +-sub-location of cells
  |        +-artificial+-cell line

+-substance+-compound+-organic+-amino+-protein+-protein family or group
  |       |       |       +-protein complex
  |       |       |       +-individual protein molecule
  |       |       |       +-subunit of protein complex
  |       |       |       +-substructure of protein
  |       |       |       +-domain or region of protein
  |       |       +-peptide
  |       |       +-amino acid monomer

+-nucleic+-DNA+-DNA family or group
  |       |       +-individual DNA molecule
  |       |       +-domain or region of DNA

+-+-RNA+-RNA family or group
  |       |       +-individual RNA molecule
  |       |       +-domain or region of RNA
Expansion of GENIA Ontology

• Try to tag all NPs in some MEDLINE abstracts and find the classes that appear in abstracts but not in current ontology
• Find frequent verbs and what class of arguments they take
Expansion of GENIA Ontology

• Chemical class of substance and their substrucutres
• Sources
• Biological role, or function, of substances
• Reaction
  – Biological reaction
  – Pathway
  – Disease
• Structure themselves
• Experiment, experimental results, and researchers
• Measure
Example of Entities in Expanded

- Biological role, or function, of substances
  - receptor, inhibitor, …
- Biological reaction
  - activation, binding, inhibition, apoptosis, G2 arrest
  - pathway, signal
  - immune dysfunction, Ataxia telangiectasia (AT)
- Structure themselves
  - alpha-helix,
- Experiment, experimental results, researchers
  - our results, these studies, we
## Verbs Related to Biological Events

### Frequent Verbs in 100 MEDLINE Abstracts

<table>
<thead>
<tr>
<th>Verb</th>
<th>Count</th>
<th>Verb</th>
<th>Count</th>
<th>Verb</th>
<th>Count</th>
<th>Verb</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>be</td>
<td>255</td>
<td>involve</td>
<td>16</td>
<td>determine</td>
<td>9</td>
<td>explain</td>
<td>6</td>
</tr>
<tr>
<td>induce</td>
<td>56</td>
<td>identify</td>
<td>16</td>
<td>construct</td>
<td>9</td>
<td>exert</td>
<td>6</td>
</tr>
<tr>
<td>bind</td>
<td>50</td>
<td>act</td>
<td>15</td>
<td>associate</td>
<td>9</td>
<td>enhance</td>
<td>6</td>
</tr>
<tr>
<td>show</td>
<td>49</td>
<td>stimulate</td>
<td>14</td>
<td>reduce</td>
<td>8</td>
<td>display</td>
<td>6</td>
</tr>
<tr>
<td>suggest</td>
<td>42</td>
<td>profile</td>
<td>14</td>
<td>prevent</td>
<td>8</td>
<td>characterize</td>
<td>6</td>
</tr>
<tr>
<td>activate</td>
<td>42</td>
<td>express</td>
<td>13</td>
<td>bcate</td>
<td>8</td>
<td>participate</td>
<td>5</td>
</tr>
<tr>
<td>factor</td>
<td>36</td>
<td>affect</td>
<td>13</td>
<td>line</td>
<td>8</td>
<td>localize</td>
<td>5</td>
</tr>
<tr>
<td>demonstrate</td>
<td>35</td>
<td>type</td>
<td>12</td>
<td>differ</td>
<td>8</td>
<td>investigate</td>
<td>5</td>
</tr>
<tr>
<td>inhibit</td>
<td>26</td>
<td>report</td>
<td>12</td>
<td>trigger</td>
<td>7</td>
<td>imply</td>
<td>5</td>
</tr>
<tr>
<td>have</td>
<td>25</td>
<td>form</td>
<td>12</td>
<td>synergize</td>
<td>7</td>
<td>establish</td>
<td>5</td>
</tr>
<tr>
<td>reveal</td>
<td>21</td>
<td>contribute</td>
<td>12</td>
<td>examine</td>
<td>7</td>
<td>conclude</td>
<td>5</td>
</tr>
<tr>
<td>require</td>
<td>21</td>
<td>study</td>
<td>11</td>
<td>block</td>
<td>7</td>
<td>compare</td>
<td>5</td>
</tr>
<tr>
<td>regulate</td>
<td>21</td>
<td>observe</td>
<td>11</td>
<td>become</td>
<td>7</td>
<td>use</td>
<td>4</td>
</tr>
<tr>
<td>indicate</td>
<td>21</td>
<td>lead</td>
<td>11</td>
<td>analyze</td>
<td>7</td>
<td>transform</td>
<td>4</td>
</tr>
<tr>
<td>find</td>
<td>21</td>
<td>function</td>
<td>11</td>
<td>target</td>
<td>6</td>
<td>transfec</td>
<td>4</td>
</tr>
<tr>
<td>result</td>
<td>20</td>
<td>assay</td>
<td>11</td>
<td>signal</td>
<td>6</td>
<td>test</td>
<td>4</td>
</tr>
<tr>
<td>phy</td>
<td>19</td>
<td>appear</td>
<td>11</td>
<td>remain</td>
<td>6</td>
<td>suppress</td>
<td>4</td>
</tr>
<tr>
<td>interact</td>
<td>18</td>
<td>occur</td>
<td>10</td>
<td>produce</td>
<td>6</td>
<td>support</td>
<td>4</td>
</tr>
<tr>
<td>mediate</td>
<td>17</td>
<td>increase</td>
<td>10</td>
<td>present</td>
<td>6</td>
<td>substitute</td>
<td>4</td>
</tr>
<tr>
<td>contain</td>
<td>17</td>
<td>phosphorylate</td>
<td>9</td>
<td>possess</td>
<td>6</td>
<td>share</td>
<td>4</td>
</tr>
</tbody>
</table>
Verbs Related to Biological Events

Verbs that take biological entities as arguments

• induce
  – noun BE INDUCED BY noun
    activation of these PROTEIN was induced by PROTEIN
  – noun INDUCE noun
    PROTEIN induced the tyrosine phosphorylation

• bind
  – noun BIND TO noun
    the drugs bind to two different PROTEIN
  – noun BIND noun
    motifs previously found to bind the cellular factors
  – noun BINDING noun
    the TATA-box binding protein
  – the BINDING of noun
    the binding of PROTEIN

semantic class: substance structure source experiment fact reaction
Verbs Related to Biological Events

Verbs that take description entities

• report
  – noun REPORT that-clause  \(\text{we report here \ that PROTEIN is activated by PROTEIN}\)
  – noun REPORT noun  \(\text{we report the characterization of PROTEIN}\)
  – noun REPORT noun  \(\text{we report a novel structure of PROTEIN}\)

semantic class: substance structure source experiment fact reaction
Verbs Related to Biological Events
Verbs whose arguments depend on syntactic patterns

• show
  – noun BE SHOWN to-infinitive PROTEIN has been shown to trigger cellular PROTEIN activity
  – noun SHOW that-clause the data show that PROTEIN stimulation is also not sufficient
  – noun SHOW noun SOURCE showed a dose-dependent inhibition of PROTEIN activity

semantic class: substance source experiment fact
Verbs Related to Biological Events
Verbs that take both entities

• indicate

– noun INDICATE that-clause  the data indicate that PROTEIN is required in CELL proliferation
– noun INDICATE noun  these findings indicate an unexpected role of DNA

– noun INDICATE that-clause  the structure indicates that it represents a unique class of PROTEIN
– noun INDICATE noun  the structure indicates mechanisms for allosteric effector action

semantic class: substance structure source experiment fact reaction role
1. IE can contribute to Bio-informatics significantly.

2. However, the domains in Bio-chemistry seem more structurally rich than the domains we have dealt with so far.
   Term formation, rich ontologies, complex syntactic structures.

3. It requires substantial efforts in resource building.

4. However, those resources can contribute to other applications:
   Knowledge sharing, Intelligent IR, Knowledge discovery

   One of the crucial techniques is ATR ….
Overview of GENIA System

Corpus Module
- Markup generation / compilation
- Annotated corpus construction

Information Extraction Module
- Identify & classify terms
- Identify events

Retrieval Module
- Request enhancement
- Spawn request
- Classify documents

Interface Module
- GUI
- HTML conversion
- System integration

User
- IR Request
- Abstract
- Full Paper

Background Knowledge
- Ontology
- Markup language
- Data model

Concept Module
- BK design / construction / compilation

Database Module
- DB design / access / management
- DB construction

Document
Named-Entity
Event

Security

Corpus
- Raw(OCR)
- Text Structure
- Annotated

Database

MEDLINE