The HINOKI (檜) Project
~ Toward Text Understanding

Francis Bond*, Sanae Fujita*, Chikara Hashimoto**, Eric Nichols***, Takaaki Tanaka*

*NTT Communication science lab.  **Kyoto University, ***NAIST
HINOKI Project: Motivation

- Make computers clever
- Natural language processing based on both structural and lexical semantics.
- Ex) Information Retrieval, QA, etc…
  - ライターを雇いたい。  
    *I want to hire a writer.*
  - ライターを買いたい。
    *I want to buy a lighter.*
What’s needed? (1)
~Grammar (Jacy) & Treebank

• Using Grammar Jacy, DELPH-IN tools (LKB. PET, [incr tsdb()]).

Ex) Definition Sentences of ライター (writer)

The text ACC wrote person
What’s needed? (2)
~Dictionary (Lexeed)

- Lexeed (Seed of Lexicon)
- The most familiar words of Japanese
  - Familiarity is estimated by psychological experiments
- 28,000 words, 46,000 senses
- Covers 75% of tokens in a typical newspaper.
- Self-contained dictionary
- Has def. sentences and at least one ex. for each sense.

Ex) Definitions

ライター_2 (writer) [noun]
(2-1) a person who has written the writing. author.

ライター_1 (lighter) [noun]
(1-1) light. igniter. a device for lighting or igniting fuel, especially cigarette.
ex) He lit the cigarette with his lighter.
What’s needed? (3) ~Sensebank

We tagged some corpora using the Lexeed senses.

ライターに依頼した。 I asked a writer.

ライターを買った。 I bought a lighter.

ライターが必要だ。 I need a writer/lighter.

Definition of Lexeed

ライター_2 (writer) [noun]
(2-1) a person who has written the writing. author.

ライター_1 (lighter) [noun]
(1-1) light. igniter. a device for lighting or igniting fuel, especially cigarette.
What’s needed? (4) ~Ontology

- Extracting from RMRS of Definition sentences of Dictionary

For Japanese and English
HINOKI Project: Status

- Constructing resources and making fundamental parsers

Rich Resources
- Grammar: JACY
- Syntactic: HINOKI Treebank
- Semantics: HINOKI Sensebank
- Ontology
- Dictionary: Lexeed

Finished first round

Building / Improving
HINOKI Project: Domain

- Expanding from closed world to open domain, from hand-build to semi-auto.
HINOKI Treebank: Status

- Finished first round for dictionary
  - 75% of def. and ex. sentences have correct trees
- Now annotating newspaper articles
  - 47% (Kyoto) and 25% (Senseval 2) are parsed, of those around 50% are correct
- Need more Grammar and Lexicon development

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sentences</th>
<th>Words</th>
<th>have Correct Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexeed Def.</td>
<td>75,000</td>
<td>691,000</td>
<td>54,000</td>
</tr>
<tr>
<td>Lexeed Ex.</td>
<td>45,000</td>
<td>499,000</td>
<td>35,000</td>
</tr>
<tr>
<td>News (Kyoto Univ.)</td>
<td>38,000</td>
<td>969,000</td>
<td>*4,600</td>
</tr>
<tr>
<td>News (Senseval 2)</td>
<td>36,000</td>
<td>888,000</td>
<td>*4,400</td>
</tr>
</tbody>
</table>
HINOKI Sensebank: Status

- Biggest Japanese sensebank
  - Senseval2 (Japanese): 150,000
  - SemCor (English): 230,000

In total, 819,000 Corpus Words (No.)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Words (No.)</th>
<th>Tagged</th>
<th>agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexeed Def.</td>
<td>691,000</td>
<td>199,000</td>
<td>78.7</td>
</tr>
<tr>
<td>Lexeed Ex.</td>
<td>499,000</td>
<td>127,000</td>
<td>82.0</td>
</tr>
<tr>
<td>News (Kyoto)</td>
<td>969,000</td>
<td>224,000</td>
<td>83.2</td>
</tr>
<tr>
<td>News(Senseval2)</td>
<td>888,000</td>
<td>269,000</td>
<td>83.3</td>
</tr>
</tbody>
</table>
HINOKI Ontology: Status

- Created from several dictionaries
- Compared with existing thesaurus
  - Consistency: ratio of extracted relations which are consistent with existing thesaurus (Goi-taikei, WordNet)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexeed</td>
<td>Ja</td>
<td>28,000</td>
<td>46,000</td>
<td>77,000</td>
<td>57%*</td>
</tr>
<tr>
<td>Iwanami(RWCP)</td>
<td>Ja</td>
<td>60,000</td>
<td>86,000</td>
<td>142,000</td>
<td>60%</td>
</tr>
<tr>
<td>GCIDE</td>
<td>En</td>
<td>130,000</td>
<td>171,000</td>
<td>78,000</td>
<td>36%**</td>
</tr>
</tbody>
</table>

* : Evaluated by hand: 89%
**: used only hypernym 100,000 sentences (12,000 words, 36,000 senses)
Future Work

- Automate analysis
- Word sense tagger (prototype is made)
- Parse ranking
- Combining them
- Expand the system and resources
- Expand the coverage of the Grammar and Dictionary
- Enable the system to treat unknown words