Grammar engineering work at U-Tokyo

Yusuke Miyao
University of Tokyo
Project overview

Grammar engineering

Parsing technologies

Enju  HPSG-based English parser

Resource development for bio domain

Treebank
Named entities
Biological events...

NER, synonym extraction, etc.

Machine learning

Machine translation

Bio IR/IR

GENIA

NER, synonym extraction, etc.
MEDIE

- A search engine for biomedical papers
- Semantic search: specify subject/predicate/object
- Synonym expansion for protein names, event expressions, etc.

In conclusion, we demonstrated for the first time that activation of phosphatidylinositol-3-kinase (PI-3K)-Akt and ERK2 pathways significantly contributed to cardioprotective effects of a Ca (2+) antagonist and a beta-adrenergic receptor blocker.

Recently, we found that all-trans retinoic acid (atRA) triggers the activation of extracellular-signal-regulated kinase 2 (ERK2), which phosphorylates TR2 and stimulates its partitioning to promyelocytic leukemia (PML) nuclear bodies, thereby converting the activator function of TR2 into repression (Gupta et al. 2008; Park et al. 2007).

http://www-tsujii.is.s.u-tokyo.ac.jp/medie/
What is Enju?

- An English parser based on the HPSG theory [Pollard and Sag, 1994]
- Fast, robust, accurate analysis of phrase structures and predicate argument structures
- HPSG grammar (lexicon & grammar rules) + probabilistic model for disambiguation
  - An HPSG treebank is constructed from Penn Treebank
  - A lexicon and a probabilistic model are obtained from the HPSG treebank
Topic of this talk

- Motivation: difficulty in the development of wide-coverage linguistic grammars
- Our solution: Corpus-oriented development of an HPSG grammar
  - The principal aim of grammar development is treebank construction
  - Penn Treebank is converted into an HPSG treebank
  - A lexicon is extracted from the HPSG treebank
Background: HPSG

• HPSG is a syntactic theory to explain generic regularities that underlie phrase structures, lexicons, and semantics [Pollard & Sag 1994]

• Two components of HPSG:
  – Lexical entries represent word-specific constraints
  – Grammar rules express generic grammatical regularities
Background: HPSG parsing

- Lexical entries determine syntactic/semantic constraints of words

Lexical entries

- [HEAD noun
  SUBJ <>
  COMPS <>]
- [HEAD verb
  SUBJ <HEAD noun>
  COMPS <HEAD noun>]
- [HEAD noun
  SUBJ <>
  COMPS <>]

John saw Mary
Background: HPSG parsing

- Grammar rules determine generic constraints of grammar (not limited to construction rules)

```
[HEAD noun
  SUBJ <>
  COMPS <>]  [HEAD verb
  SUBJ <HEAD noun>
  COMPS <HEAD noun>]
```

John    saw    Mary
Background: HPSG parsing

- Grammar rules determine generic constraints of grammar (not limited to construction rules)
Background: HPSG parsing

- Grammar rule applications produce syntactic/semantic structures of sentences

```
  HEAD verb
     SUBJ <>
     COMPS <>

  HEAD noun
     SUBJ <>
     COMPS <>

  HEAD verb
     SUBJ <HEAD noun>
     COMPS <>

  HEAD noun
     SUBJ <>
     COMPS <>
```

John saw Mary

11
Requirements

• For HPSG parsing, we require:
  – Grammar rules
  – Lexical entries
  – Treebank
    • For statistical modeling
    • For grammar testing

What is the fastest way to the development of these three resources?
If we have an HPSG treebank, we can collect lexical entries from terminal nodes
Our approach

1. Develop grammar rules and an HPSG treebank
2. Collect lexical entries from the HPSG treebank
How to make an HPSG treebank?

• Convert Penn Treebank into HPSG-conformant structures
• Grammar development = restructuring a treebank in conformity with HPSG grammar rules
Overview of grammar development

1. Treebank conversion

2. Grammar rule application

3. Lexical entry collection
1. Treebank conversion

- Modify constituent structures
- Add feature structures
Example: auxiliary/control verbs

- Auxiliary/control verbs are annotated as taking unsaturated constituents
Example: object relative

- SLASH represents moved constituents
- REL represents relative-antecedent relations
2. Grammar rule application

• Grammar rules are applied to HPSG-style parse trees
  – Grammar rule application fails when a parse tree contains errors/inconsistencies
  – Unspecified feature values are filled

• Resulting parse trees are assured to satisfy constraints of the HPSG theory
Example

• “NL is officially making the offer”
Example

• “NL is officially making the offer”
3. Lexical entry collection

• Collect terminal nodes of HPSG parse trees
• Assign predicate argument structures
Collecting terminal nodes

- Terminal nodes of HPSG parse trees are instances of lexical entries

```
[HEAD verb
  SUBJ <>
  COMPS <>
]
```

```
[HEAD noun
  SUBJ <>
  COMPS <>
]
```

```
NL
```

```
[HEAD verb
  SUBJ <>
  COMPS <>
]
```

```
[HEAD verb
  SUBJ <>
  COMPS <>
]
```

```
[HEAD verb
  SUBJ <>
  COMPS <>
]
```

```
[HEAD verb
  SUBJ <>
  COMPS <>
]
```

```
[HEAD verb
  SUBJ <>
  COMPS <>
]
```

```
[HEAD verb
  SUBJ <>
  COMPS <>
]
```

```
[HEAD verb
  SUBJ <>
  COMPS <>
]
```

```
[HEAD noun
  SUBJ <>
  COMPS <>
]
```

is

officially

making

the offer
Assigning predicate argument structures

- Create mappings from syntactic arguments into semantic arguments

Ex. lexical entry for “make”
Evaluation

• Data
  – Lexicon extraction and training of disambiguation models: Section 02-21 (39,832 sentences)
  – Test: Section 22 (1,700 sentences)

• Results
  – Coverage: 99.7%
  – Accuracy of predicate-argument relations
    • Precision: 88.01%
    • Recall: 87.70%
Multilingual grammar development

• English HPSG
  – Penn Treebank
  – Domain adaptation using GENIA and Brown

• Chinese HPSG
  – Penn Chinese Treebank

• Japanese CCG
  – Kyoto Text Corpus
Summary

• Corpus-oriented development of an HPSG grammar is presented
• A wide-coverage HPSG lexicon is obtained, and accurate parsing is achieved
• Multilingual grammar development is underway
  – Chinese HPSG from Penn Chinese Treebank
  – Japanese CCG from Kyoto Text Corpus