A Robust Unification-based Parsing Framework

Ai Azuma

Computational Linguistics Lab.
Nara Institute of Science and Technology

Background

o Analyzing informal texts

Formal documents
- Newspapers
- Books
- Dictionaries

Informal documents
- Personal home pages
- BBSes
- Blogs

Natural Language Interfaces
- Input to QA system

Error Recovery Feature

o Error recovery feature
- Detect what kind of errors occur in input
- Recover the errors

o Applications of this feature
- Aid for word processing
- Educational support system for second language learners

Difficulty of Robust Parsing

o Why it is difficult to analyze informal documents?
- Even with the best effort, not all sentences can be covered
- Grammatical rules for exceptional phenomena can over-generate, causing a drop in precision

---

Related Work (1/2)

o Individual treatments for each errors
- Rephrasing [Sagawa et. al., 1994]
- Particle ellipsis and inversion [Yamamoto et. al. 1992]

Related Work (2/2)

o General approach for errors
- Multi-staged approach
  - Partial parsing [McDonald, 1992]
  - Two-staged analysis with meta-rules [Weischedel et. al., 1983]
  - Bottom-up parsing with top-down error detection [Kato, 1992]
- Unified approach
  - Preference semantics [Fass et. al., 1983]
  - Abduction [Hobbs et. al., 1993]

Goal of this research

- Robust parsing framework with error recovery feature
  - Unified treatments of errors – view errors as unification failures
  - Coverage of error recovery is controlled by "cost"

Coverage of error recovery module

System's grammar

"Real" grammar

Approach
- Error Recovery Module

- Typical parsing framework with additional error recovery module
  - For well-formed inputs, the system behaves the same as typical system, and the module is never used
  - For ill-formed inputs, the system uses the module in order to output the most likely analysis

Cost and Preference of Error Recovery Methods

- Cost determines which error recovery methods are preferred

[   ]   [   ]

Which is preferred?

[   ][   ]

Determining Cost

- Corpus-based approach
  - Based on distribution of errors in a corpus, cost is divided up among errors

[   ]   [   ]

[   ][   ]

Determining Cost

- Corpus-based approach
  - n distribution of errors in a ost is divided up among

If the lack of WO-particle is more frequent than the lack of GA in the verb " " in a corpus, then…

[   ]   [   ]

This is preferred.

[   ][   ]

Case Study
– In a Spontaneous Corpus

- Zero pronoun ellipsis
  - Detect lack of cases
  - Fill cases with a dummy noun (zero)

[   ][   ]

Case Study
– In a Spontaneous Corpus

- Hesitation
  - Can be simply treated by registering fillers in dictionaries
  - May be confused with words of other categories

1

Adnominal or filler?

Case Study
– In a Spontaneous Corpus

- Rephrasing
  - Patterns may widely vary
  - Difficult to identify rephrased target

Case Study
– In a Spontaneous Corpus

- Repetition

Case Study – In a Spontaneous Corpus

- Particle Ellipsis

[ ] [ ]

object   subject   Ambiguity due to lack of case marker

Similar techniques could be applied

object   subject   The same kind of ambiguity occurs

Future Direction

- Categorizing errors, especially rephrasing in spontaneous corpuses
- Determining cost

Categorizing Errors

- Categorizing rephrasing errors

Phonological rephrasing

CSL

Syntactical rephrasing

Semantic rephrasing

Determining cost

- Count errors in large corpus and divide up costs among errors
- Current research focus is to find suitable framework
  - Unsupervised learning approach
  - Data mining approach

Summary

- Background: Needs to parse informal texts
- Research goal:
  - To build a robust parsing framework
  - Error modification
- Approach: Add an error recovery module to an ordinal parsing framework
- Future Direction:
  - Categorize and annotate errors
  - Find suitable framework to determine cost