Ambiguity & Ambiguity Resolution

11-731 Machine Translation

Teruko Mitamura
Language Technologies Institute
Carnegie Mellon University

Ambiguity

defense
strong defense
final
conference
good line
Penguins

Lexical Ambiguity

• **Category Ambiguity:** a word is assigned to more than one grammatical category (e.g. noun, verb, or adj)
  
  N/V
  N/V/Adj/Adv “last”
  “Gas pump prices rose last time oil stocks fell.”

Lexical Ambiguity (2)

• **Homography:** two or more ‘words’ with quite different meanings which have the same spelling (e.g. bank, ball)

• **Polysemy:** a range of meanings related in some way to each other. (e.g. metaphorical extension, mouth of a river, branch of a bank)

Lexical Ambiguity (3)

• **Transfer Ambiguity:** a single source language word can be translated by a number of different target language words.

Real Structural Ambiguities VS System Ambiguities

• ‘Real’ ambiguities: a human might find several interpretations
• ‘System’ ambiguities: the grammar makes distinctions. A human reader would not necessarily recognize.
Real Structural Ambiguity

“Flying planes can be dangerous.”
It can be dangerous to fly planes.
Planes which are flying can be dangerous.

“Time flies like an arrow.”
The passage of time is as quick as an arrow.
A species of flies called ‘time flies’ enjoy an arrow.
Measure the speed of flies like you would arrow.

“The man saw the girl with the telescope.”

Attachment Ambiguities

• Prepositional phrase attachment
• Relative clause
• Noun-noun compounding
• Adjective modifiers

Disambiguation

• POS Tagging
  – Stochastic Tagging: probabilities on tag sequences
  – Rule-Based Tagging
• Issues in Tagger Design
  – Tokenization — a number of strings are difficult to classify: she'd, shouldn't, intra- and extra-mural, inasmuch as, etc.
  – Human Assignment — a number of grey areas (up to 3.5% disagreement)
  – Tagset Design — tagsets range from a few tens of tags up to 197 tags (London-Lund Corpus of Spoken English) For automatically induced tagsets, the categories may be too broad for practical applications

Word Sense Disambiguation

• Selectional Restrictions
  “The crane flew over the plain.” (bird)
  “The builder operated the crane.” (machine)
• Frames and Semantic Distance
  – Use of knowledge representation (is-a, has_part, etc.)
  – The general assumption of the preferred sense of a word: semantically closest to the sense of the word(s) which combines with.

Frames

```
physical_object

Wing  animal  machine

Has_part

bird  crane-2

Is-a

```

Word Sense Disambiguation (2)

• Corpus-Based Techniques
  – Determine the most frequent sense of a word from a sense-tagged corpus.
    • This simple approach yields around 70% accurate tagging for general English text.
    • Difficulties of manually sense-tagging a training corpus
    • Data sparseness
  – Automatic Sense-Tagging
    • Bootstrapping methods — improving accuracy incrementally
    • Use of bilingual corpus
Structural Disambiguation

- Psycholinguistic Preferences
- Minimal Attachment (Frazier and Foder, 1978)
  - A constituent tends to attach so as to involve the fewest additional syntactic nodes
- Right Association (Kimball, 1973)
  - A constituent tends to attach to another constituent immediately to its right

These two principles make opposite predictions.

Example

Mary hid the photo in the drawer.

Example (cont.)

Mary hid the photo in the drawer.

Structural Disambiguation (2)

- Probabilistic Context-Free Grammars
  - NP -> Pron 0.45
  - NP -> Det N 0.35
  - NP -> N 0.20
- POS labels may be too broad (e.g. no subcat information)
- Do not take lexical preferences into account
- Lack of information on grammatical relations

Coping with Ambiguity in KANT

- The strategies and techniques used to cope with ambiguity in the KANT Knowledge-based Machine Translation system.
- What is ambiguity in KANT?
  - More than one interlingua representation of a word, phrase, or a sentence.
- Implication for MT
  - Reduction in quality if ambiguity exists.
  - One way to improve the quality of an MT system is to reduce the amount of ambiguity.
Constraining the Source Text

- Domain Lexicon from Corpus
  - Single-word lexical entries: ~10,000
  - Technical phrase entries: ~80,000
  - Domain-specific part of speech assignment
  - Usage definition and examples from corpus
- Grammar
  - Define the set of possible grammatical structures
  - Use of SGML tags to disambiguate words and structures
  - Structural disambiguation using semantic domain model (DM)

Constraining the Source Text (2)

- Disambiguation by the author
  - Lexical Disambiguation (~100 ambiguous lexical items)
  - Structural Disambiguation for PPs (not reduced by DM)

Example of Ambiguous Lexical Entries:

```
((:ROOT "apply")
 (:POS V)
 (:CONCEPT *A-APPLY-1)
 (:CLASS AGENT+THEME)
 (:NOTE
 (:SENSE "Meaning 1: to put into contact, or to a special use"
  "When installing a new filter element, apply a small amount of
  clean engine oil to the rubber seal.")
 (:ACTION :HYD :AMBIG :AMBIGTRUE)
 (:FREQUENCY 2019 3440))
```

Example of Ambiguous Lexical Entries:

```
((:ROOT "apply")
 (:POS V)
 (:CONCEPT *A-APPLY-2)
 (:SYN-FEATURES (VALENCY TRANS INTRANS))
 (:CLASS THEME/AGENT+THEME)
 (:NOTE
 (:SENSE "Meaning 2: to put into action"
  "Push the brake pedal down
  in order to apply the brakes.")
 (:ACTION :HYD :AMBIG :AMBIGTRUE))
```

Examples of SGML Tags

```
<para>
 <callout>1</callout> Verify button.
 <callout>2</callout> Reset button.
 <callout>3</callout> LED overspeed light.
 <callout>4</callout> Seal screw plug for overspeed adjustment.
 <callout>5</callout> Seal screw plug for crank terminate adjustment.
 <callout>6</callout> Seal screw plug for oil step.
</para>
```

Semantic Domain Model

- Used for disambiguation of syntactic attachment (e.g. PP-attachment, Adj-Noun attachment)
  - Concepts are arranged in an inheritance hierarchy
  - DM is consulted at the earliest possible stage during parsing
  - Grammar rule succeeds only if the attachment is semantically licensed
  - Efficient Run-time use of semantic restrictors
    - indexed by head concept
    - all inheritance is performed off-line
    - stored in a space-efficient structure-shared manner
Lexical Disambiguation by the Author

• Ask author to choose the right meaning and part of speech

  "Do not apply any retaining compounds or any anti-seize compounds to the cylinder head surfaces."

  apply-1 - to put into contact, or to a special use
  apply-2 - to put into action

  "Do not apply "<CTE means text='apply' val='*A-APPLY-1'> any retaining compounds or any anti-seize compound to the cylinder head surfaces."

Structural Disambiguation by the Author

Ask author to choose the right structural attachment for PPs

  "Install the bridge, the thrust bearing, the washer and the nut from tooling <callout> A </callout> over the stud."

  install over
  nut over
  tooling over

  "Install the bridge, the thrust bearing, the washer and the nut from tooling <callout> A </callout> over "<CTE attach head='install' modi='over'> the stud."

Disambiguation in KANT

850 sentences were tested.

<table>
<thead>
<tr>
<th>Test</th>
<th>Lex</th>
<th>Gra</th>
<th>N-N</th>
<th>DM</th>
<th>#Parse</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gen</td>
<td>Gen</td>
<td>Yes</td>
<td>No</td>
<td>27.0</td>
</tr>
<tr>
<td>2</td>
<td>Gen</td>
<td>Gen</td>
<td>No</td>
<td>No</td>
<td>10.2</td>
</tr>
<tr>
<td>3</td>
<td>Gen</td>
<td>Con</td>
<td>Yes</td>
<td>No</td>
<td>8.4</td>
</tr>
<tr>
<td>4</td>
<td>Con</td>
<td>Gen</td>
<td>Yes</td>
<td>No</td>
<td>1.7</td>
</tr>
<tr>
<td>5</td>
<td>Con</td>
<td>Gen</td>
<td>No</td>
<td>No</td>
<td>1.6</td>
</tr>
<tr>
<td>6</td>
<td>Con</td>
<td>Con</td>
<td>No</td>
<td>Yes</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Disambiguation Results (750 sentences)

<table>
<thead>
<tr>
<th>Test</th>
<th>Lex</th>
<th>Gra</th>
<th>N-N</th>
<th>DM</th>
<th>Disambig</th>
<th>#parse</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Con</td>
<td>Con</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>1.04</td>
</tr>
</tbody>
</table>

• 95.6% of sentences have a single interlingua representation
• Small amount of residual ambiguity

Summary

• Constraining the lexicon achieved the largest reduction of ambiguity
• Semantic knowledge bases disambiguate structural attachment in KANT
• Generation Strategies:
  – Generate TL sentences from all of the interlingua representations
  – Generation heuristics are used to pick one TL sentence, when >1 IR.

Summary (2)

• Building and maintaining semantic knowledge bases requires significant effort
• Interactive disambiguation also requires significant effort from authors
• Proposed solution
Reasons for Improving Automatic Disambiguation

• PP attachment ambiguity averages about one PP per sentence in heavy equipment documentation
• Unresolved ambiguity leads to higher translation cost (one interpretation is picked randomly, leading to reduced translation quality)

Reasons for Improving Automatic Disambiguation (2)

• Interactive disambiguation leads to higher authoring costs
• Authors make errors, sometimes choosing the wrong attachment

Steps to analyze each sentence in KANT

1. Morphological analysis is performed
2. The set of possible lexical entries is retrieved
3. A unification grammar is used to build an f-structure
4. When >1 possible f-structure, a semantic domain model is used to disambiguate

Steps to analyze each sentence in KANT (2)

5. If ambiguity still exists after semantic disambiguation, interactive disambiguation is triggered

Structural Disambiguation in KANT

• Semantic Domain Model contains attachment preferences in the form of triples (<head> <semantic-role> <filler>)
(*A-LIFT INSTRUMENT *O-HOIST) “Lift the engine with a hoist.”

Semantic Domain Model

1. Each PP attachment is checked against the triples
2. Assignment of a score
   – match a triple exactly: 0
   – match a triple under IS-A inheritance on the head or filler: 1
   – match a triple under IS-A inheritance on both head and filler: 2
Semantic Domain Model (2)

- The attachment scores are summed
- The set of f-structures with the lowest score is retained
- Result: the f-structures which most closely match the specific domain knowledge are preferred

Interactive Disambiguation

- For truly ambiguous sentences in the domain
- For sentences where no semantic knowledge was applied
- Author chooses the attachment site
- Author’s choice is inserted into the source text using an SGML processing instruction

Experimental Corpus

- Test corpus: 12,000 sentences
- PP attachment: 11,607 instances
- Disambiguation by Domain Model:
  – 8,209 PP attachments (71% of the total)
  – 9,254 sentences (77% of the total)
- Interactive Disambiguation:
  – 3,398 PP attachments (29% of the total)
  – 2,748 sentences (23% of the total)

Goal of Experiment

- Improve Automatic Disambiguation
- Eliminate Interactive Disambiguation
- How?
  – Introduction of Automatic Attachment Heuristics

Automatic Attachment Heuristics

1. Disambiguation using Semantic Domain Model (existing)
2. Syntactic VERB+Prep Attachment Patterns
3. Other Syntactic Attachment Patterns
4. Default (Local) Attachment

Syntactic VERB+Prep Attachment Patterns

- USE + something + FOR
  “Do not use a chain for pulling.”
- INSTALL + something + OVER
  “Install the plastic cup over the bolt.”
- PROVIDE + something + FOR
  “This force provides the power for the brake application.”
Other Syntactic Attachment Patterns

- VERB + something + BY + VERB -ing
  “Release the pressure by loosening the filler cap.”
- VERB + something + FOR + VERB -ing
  “Use a scraper to rip the surface for loading.”

Default (Local) Attachment

- “Raise the air pressure in the line to the cylinders.”
- “The check valve allows steering with a dead engine.”

Experiment (1)

- Compared Interactive Disambiguation and Automatic Attachment Heuristics
- 2748 sentences from the corpus, where Interactive Disambiguation was used originally
  - 65 Verb+Prep Patterns were implemented
  - 4 Other Patterns were implemented

Results

The KANT Spanish translation output results

<table>
<thead>
<tr>
<th>Method Used</th>
<th>Correct Trans.</th>
<th>Incorrect Trans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive</td>
<td>2442 (89%)</td>
<td>309 (11%)</td>
</tr>
<tr>
<td>Default Attachment Only</td>
<td>2290 (83%)</td>
<td>458 (17%)</td>
</tr>
<tr>
<td>Patterns + Default</td>
<td>2449 (89%)</td>
<td>299 (11%)</td>
</tr>
</tbody>
</table>

Experiment (2)

- Test Sentences: 1445
- KANT Spanish system (July, 1999)
- Compared Interactive and Automatic Disambiguation
  - Original corpus (Interactive Disambiguation) is translated to Spanish and scored
  - Compared with scores from the same corpus translated with Automatic Disambiguation

Results (2)

- Regression Testing on 1,445 sentences
  - Total Differences: 88
  - Scores remains the same: 40
  - Improvements: 31
  - Became worse: 17
  - Total improvements: 14
Findings

- No significant difference in the quality of the Spanish translation output
- Some incorrect automatic attachments, which didn’t affect the output
- Productivity increase on the authoring side by eliminating interactive disambiguation

Conclusion

- Multiple strategies for automatic semantic and syntactic disambiguation can work
- Minimal impact on translation quality for 12,000 sentences (2,447 interactive)
- Disambiguation by Semantic Domain Model still plays an important role (3/4 of the cases) in KANT

Questions?