Interlingua MT

11-731 Machine Translation
February 7, 2007

Teruko Mitamura
Language Technologies Institute
Carnegie Mellon University

State-of-the-Art in MT (review)

• What users want:
  - General purpose (any text)
  - High quality (human level)
  - Fully automatic (no user intervention)

• We can meet any 2 of these 3 goals today, but not all three at once:
  - FA HQ: Knowledge-Based MT (KBMT)
  - FA GP: Corpus-Based (Example-Based) MT
  - GP HQ: Human-in-the-loop (efficiency tool)

Types of MT Applications:

• Assimilation:
  - Multiple source languages
  - Any style/topic
  - General purpose MT
  - No semantic analysis
  - GP FA or GP HQ

• Dissemination:
  - One source language
  - Controlled style
  - Single topic/domain
  - Special purpose MT
  - Full semantic analysis
  - FA HQ

Interlingua-based MT

• A “natural” deep Artificial Intelligence approach:
  - Analyze the source language into a language independent detailed symbolic representation of its meaning
  - Generate this meaning in the target language
  - “Interlingua”: one single meaning representation for all languages
  - Nice in theory, but extremely difficult in practice

The Interlingua KBMT approach

• With interlingua, need only N parsers/generators instead of N^2 transfer systems:

Advantages of Interlingua

• Add a new language easily
  - get all-ways translation to all previous languages by adding one module for analysis and one module for generation

• Mono-lingual development teams.

• Paraphrase
  - Generate a new source language sentence from the interlingua so that the user can confirm the meaning

• Language-independent representation
Problems of Interlingua

• “Meaning” is arbitrarily deep.
  – What level of detail do you stop at?
• If it is too simple, meaning will be lost in translation.
• If it is too complex, analysis and generation will be too difficult.
• Should be applicable to all languages
  – how do we ensure that?
• Human development time.

Problems of Interlingua (cont.)

• Difficulties of defining an interlingua
• Analysis & generation have to be strictly separated (no pair-specific transfer rule)
• Interlingua has to include all the information that might be required during the generation

Knowledge-based MT (KBMT)

• Build representations in which the content goes beyond what is linguistically implied, to contain real-world knowledge
  – “The man saw the horse with the telescope.”
  – “The man saw the girl with red hair.”

Three Knowledge Sources in KBMT

• Syntactic Grammars
  – Language dependent
  – Domain independent
  – Human-readable notation (LFG notation)
• Concept Dictionaries (Domain Model)
  – Language independent
  – Domain dependent
  – Human-readable notation (Semantic case frames)
• Mapping Rules
  – Language dependent
  – Domain dependent
  – Human-readable notation (Frame-based formalism)

KBMT at CMU

• 1986 – Center for Machine Translation (CMT) is founded
• 1986 – CMT Semsyn Demo
• 1987 – KBMT-89 begins
• 1989 – KBMT-89 ends
• 1990 – KANT prototype
• 1991 – KANT/Catalyst proof-of-concept
• 1992 – KANT/Catalyst development begins

KBMT-89 project

• SL: English and Japanese
• TL: English and Japanese
• Domain: Personal computer installation and maintenance manuals
• System: A distributed, coarsely parallel system
• Static knowledge sources
  – Ontology (domain model) of about 1,500 concepts
  – Analysis and generation lexicons (900 lexical units) and grammars
  – Syntax-to-semantics mapping rules
Issues in KBMT

• Acquiring domain knowledge is very expensive
  – Deciding which concepts to include
  – How to relate concepts to each other
  – What properties to associate with each concept
• Deciding on ontology granularity
  – Concepts map one-to-one to lexical senses
  – A small number of conceptual primitives (e.g. LCS)

KBMT/Interlingua Examples

• Semsyn: The Doctor-Patient Domain (1986)
• Concept Dictionaries (400,000) by EDR in Japan
• KANT Interlingua (1991)
• Lexical Conceptual Structures (LCS) (1993-present)
• KANTOO Interlingua (1997-present)
• Examples from Interlingua Workshop (2004)

Lexical Conceptual Structure (LCS)

• UNITRAN system (Dorr 1993)
  – English, German, Spanish
  – LCS (Jackendoff 1983, 1990) is the basis for interlingua representation

Types, Primitives, Fields

• Types: the kinds of entities
  – Event, State, Position, Path, Thing, Property, Location, Time, Manner, Intensifier, Purpose
• Types are specialized into Primitives
  – GO, STAY, BE, TO, etc.
• Primitives are further specialized by a field indicator
  – Locational, Possessional, Identificational, Temporal, Circumstantial, Existential, Perceptional, Intentional, Instrumental

LCS Example

Mary died.

{Events GOIdent
  ([Thing Mary],
  [Position TOWARDIdent
    ([Position ATIdent ([Thing Mary],
      [Property DEAD]))]))

LCS Example (2)

John killed Mary.

{Event CAUSE
  ([Thing John],
  [Events GOIdent
    ([Thing Mary],
     [Position TOWARDIdent
      ([Position ATIdent ([Thing Mary],
        [Property DEAD]))]))}
KANTOO Interlingua

- An Interlingua Frame (IF) is a recursive structure that represents a semantic concept.
- An IF consists of a head and a number of slots of different types.


KANTOO MT Modules

Disambiguation in KANTOO

- Automatic Disambiguation
  - Use of domain model to disambiguate
  - Heuristics (domain preferences)
- Interactive Disambiguation
  - Ask the author to choose
  - Annotate the input (SGML)

Disambiguation in KANTOO (2)

- Proper choice of attachment can be based on meaning
  - "The man saw the boy with the telescope."
  - "The man saw the boy with the dog."
- Attachment preferences can be used to pick one reading (even if both are syntactically possible)

Disambiguation in KANTOO (3)

- 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."
Word Sense Disambiguation

- **INPUT:**
  1. Turn the truck to the right.
  2. The deposits turn into sludge.

- **Definitions:**
  *A-TURN-1: “to cause to rotate about an axis”
  *A-TURN-2: “to change form or state”

Word Sense Disambiguation (2)

- **Domain Model Triples:**
  *A-TURN-1: (*TURN-1 (*O-RIGHT-1))
  *A-TURN-2: (*TURN-2 (*O-SLUDGE))

Word Sense Disambiguation (3)

INTERLINGUA for *A-TURN-1:

Turn the truck to the right.

*TURN-1

(argument-class agent+theme) (mood imperative) (punctuation period) (tense present)

(q-modifier

(*Q-oriented_TO

(case (*K-TO))

(object (*O-RIGHT-1

(number (:OR mass singular))

(reference definite))

(role oriented)))

(theme (*O-TRUCK (number singular) (reference definite))))

Word Sense Disambiguation (4)

INTERLINGUA for *A-TURN-2:

The deposits turn into sludge.

*TURN-2

(argument-class theme) (mood declarative) (punctuation period) (tense present)

(q-modifier

(*Q-result_INTO

(case (*K-INTO))

(object (*O-SLUDGE

(number mass) (reference no-reference))

(role result))

(theme (*O-DEPOSIT (number plural) (reference definite))))

Word Sense Disambiguation (5)

INPUT1: Turn the truck to the right.
OUTPUT1: Haga girar el camión a la derecha.
GLOSS: MAKE TURN THE TRUCK TO THE RIGHT

INPUT2: The deposits turn into sludge.
OUTPUT2: Los depósitos se convierten en sedimento.
GLOSS: THE DEPOSITS ARE TURNED INTO SLUDGE

KANTOO IL design

- Head Concepts
- Semantic Roles and Features
- Mapping Lexemes to Concepts
- Feature-Value Slots
- Anaphora Resolution
- Prepositional Phrases
### Head Concepts

<table>
<thead>
<tr>
<th>Concept Prefix</th>
<th>Denotation</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A</em></td>
<td>Action (verbal) e.g. <em>A-MAKE-SURE</em></td>
</tr>
<tr>
<td><em>O</em></td>
<td>Object (nominal) e.g. <em>O-FAX-MACHINE</em></td>
</tr>
<tr>
<td><em>M</em></td>
<td>Manner (adverbial) e.g. <em>M-THOROUGHLY</em></td>
</tr>
<tr>
<td><em>P</em></td>
<td>Property (adjectival) e.g. <em>P-MULTI-COLORED</em></td>
</tr>
<tr>
<td><em>K</em></td>
<td>(preposition) e.g. <em>K-ABOVE</em></td>
</tr>
<tr>
<td><em>INT</em></td>
<td>Intensifier e.g. <em>INT-VERY</em></td>
</tr>
<tr>
<td><em>CONJ</em></td>
<td>Conjunction e.g. <em>CONJ-THAT</em></td>
</tr>
<tr>
<td><em>QUANT</em></td>
<td>Quantifier e.g. <em>QUANT-SOME</em></td>
</tr>
<tr>
<td><em>PROP</em></td>
<td>Proper (proper nouns) e.g. <em>PROP-VERDANA</em></td>
</tr>
<tr>
<td><em>SYM</em></td>
<td>Symbol (e.g. typographic symbols) e.g. <em>SYM-EXCLAMATION-POINT</em></td>
</tr>
<tr>
<td><em>U</em></td>
<td>Unit (measurement) e.g. <em>U-AMPERE</em></td>
</tr>
<tr>
<td><em>G</em></td>
<td>Grammatical e.g. <em>G-GAPPED-ARGUMENT</em></td>
</tr>
<tr>
<td><em>Q</em></td>
<td>Role pointer (preposition) e.g. <em>Q-goal_INTO</em></td>
</tr>
<tr>
<td><em>S</em></td>
<td>Structured (complex structure, tagged) e.g. <em>S-NOT-AVAILABLE</em></td>
</tr>
<tr>
<td><em>SP</em></td>
<td>Special (domain-specific phrase) e.g. <em>SP-NOT-AVAILABLE</em></td>
</tr>
</tbody>
</table>

### Semantic Roles and Features

- **Semantic Role** is a slot that is filled with an embedded interlingua.
- **Features** contain atomic values from finite set of possible values:
  - Boolean Features vs. Non-Boolean Features e.g. (negation +), (tense present)
  - Role Pointer contains a reference to a semantic role that appears elsewhere in the IL
    e.g. (topic-role theme)
  - Gapped Information e.g. agent-less passive sentence

### Semantic Analysis Example

"The manual was printed."

(*A-PRINT (agent (*G-GAPPED-ARGUMENT (gapped +)))
 (argument-class agent+theme)
 (mood declarative)
 (tense past)
 (theme (*O-MANUAL (number singular) (reference definite))
 (topic-role theme))

### Mapping Lexemes to Concepts

- Lexical entries are not stored directly in the KANTOO IL.
- For SL analysis, the lexical mapping to concepts is stored in the SL lexicon.
- In general, KANTOO concept names are derived from the English lexical units.

### Anaphora Resolution

- Antecedent slot is used to refer to the pronominal antecedent in *G-PRONOUN* frame.

*Use the printer, if it is clean.*

(theme (*G-PRONOUN (antecedent (*G-PRONOUN (antecedent (*O-PRINTER))
 (number singular) (person third) (reference definite)))))

---

Anaphora Resolution

- Antecedent slot is used to refer to the pronominal antecedent in *G-PRONOUN* frame.

*Use the printer, if it is clean.*

(theme (*G-PRONOUN (antecedent (*G-PRONOUN (antecedent (*O-PRINTER))
 (number singular) (person third) (reference definite))))
Prepositional Phrases

- Concepts are headed by *Q-concept (semantic role + preposition)
- *Q-concept frame contains:
  1. Case, whose value is a *K-preposition concept
  2. Role – semantic roles expressed by the preposition
  3. Object – object of the preposition
- There are about 75 semantic roles for preposition

Insert the tray into the printer.

(*A-INSERT (argument-class agent+theme) (mood imperative) (punctuation period) (q-modifier (*Q-goal_INTO (case (*K-INTO)) (object (*O-PRINTER (number singular) (reference definite)) (role goal))) (tense present) (theme (*O-TRAY (number plural) (reference definite)))

Recent Interlingua Activities

- 7th Interlingua Workshop (AMTA 2004)
  http://www1.cs.columbia.edu/~habash/AMTA04-WKSHP/AMTA04-IL-WKSHP.html#finalprogram
- Interlingual Annotation of Multilingual Text Corpora (IAMTC)
  http://aitc.aiicnet.org/nsf/iamtc/

Summary of IL Design Issues

- Interlingua structures are nested or indexed
- There is a distinction between head concept, semantic roles and feature-value slots (or not)
- Complete lexical information is stored in the IL (or not)
- Concept granularity

Summary of IL Design Issues (2)

- The extent to which grammatical information is represented in IL
- Human Readable (or not)
- Representation of gapped constituents (explicitly listed or indexed)
- Representation of pronominal antecedents in IL

Questions?