Linguistic Knowledge (cont.)

11-731 Machine Translation
February 26, 2007

Teruko Mitamura
Language Technologies Institute
Carnegie Mellon University

Analysis Knowledge Sources

- In general, 'state of the art' MT systems are not based on one single linguistic theory.
- MT requires coverage of all the peculiarities of actual usage in a domain.
- Contrastive analysis of languages is required for MT.
- MT researchers have been obliged to take much more pragmatic attitudes than theoretical researchers.

Analysis/Generation Modules

1. Lexicon
2. Morphology
3. Grammar
4. Interpretation rules/Mapping rules
5. Interlingua
6. Domain Model

Design Decisions in Lexicon Development (1)

- What is a unit of lexicon entry?
  - Root form vs. Full form
  - Inflectional morphology (e.g. -ing, -ed)
  - Derivational morphology (e.g. -ly, -ment)
  - Single word entry vs. Phrasal entry (e.g. factory installed hose)
  - POS (e.g. verb, adjective, adverb)
  - Subcategory (e.g. intransitive, transitive)
  - Morphological class (e.g. Japanese: 5-dan, 1-dan verbs)

Lexical Variation Issues

- Spelling
  - Nonstandard spelling: nit vs. night
  - British spelling: colour vs. color
- Acronyms
  - PM (permanent magnet)
  - PM (phase modulation)
- Abbreviations
  - p.m. (post meridiem)
  - no. (number)
  - N.O. (normally open)
- Capitalization
  - all-caps: MIL (military specification)
  - nonstandard: BrakeSaver
- Hyphenation (e.g. de-energized, make-up valve)
- Use of Slash (e.g. check/relief valve)
Design Decisions in Lexicon Development (2)

- What kind of morphological and/or syntactic information is stored in the lexicon?
- Meaning definition included with lexical items?
- Any link(s) to semantic concept(s) for lexical items?
- Translation(s) for target language(s)?
- Usage examples for lexical items?
- Any other information for developers? (e.g. comments, frequency, update info)

Design Decisions in Morphology Development

- What level of morphological processing is needed?
  - Depends on the lexicon design
  - Inflectional processes
  - Derivational processes
  - Compounding
  - Also, depends on the grammar design

Japanese Morphology Example

Japanese morphological rules can be written in the same formalism as syntactic rules
input string: tabe-sase-rare-ta
  eat-caus-pass-past

Japanese Examples (2)

<table>
<thead>
<tr>
<th>Japanese Examples (2)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tabeta</td>
<td>eat-PAST</td>
</tr>
<tr>
<td>tabesaseta</td>
<td>eat-CAUS-PAST</td>
</tr>
<tr>
<td>taberasaara</td>
<td>eat-PASS-PAST</td>
</tr>
<tr>
<td>*taberasasaseta</td>
<td>eat-PASS-CAUS-PAST</td>
</tr>
<tr>
<td>*tabetasasense</td>
<td>eat-PASS-CAUS-PASS</td>
</tr>
<tr>
<td>*tabetetetasase</td>
<td>eat-PASS-PASS-CAUS</td>
</tr>
<tr>
<td>*tabetasasetakara</td>
<td>eat-CAUS-PASS-PASS</td>
</tr>
<tr>
<td>*tabetetetarase</td>
<td>eat-PASS-CAUS-PASS</td>
</tr>
<tr>
<td>*tabetetarere</td>
<td>eat-PASS-PASS</td>
</tr>
<tr>
<td>*tabetetasae</td>
<td>eat-PASS-CAUS</td>
</tr>
</tbody>
</table>

Design Decisions in Grammar Development

- Specify set of structures to be covered. (e.g. corpus analysis)
- For each type of structure, specify:
  - the appropriate phrase structure rules
  - the set of grammatical features (e.g., person/number/gender agreement, verb class features, etc.)
  - the grammatical functions (e.g., SUBJ, OBI, PP, etc.)
  - the feature structure (cat n) (number sg) (form pastpart)

Grammar Development

- Organize the types of rules (e.g. start rules, NP rules, VP rules, PP rules, etc.)
- Write phrase structure rules.
- Add equations to phrase structure rules.
  - Unification statement
  - Constraint equations
- Test the grammar on a test suite
Design Decisions in Interlingua Development

- **Structure of the Interlingua**
  - Concept head: Open class vs. Closed class (e.g. *A-LIFT)
  - Semantic roles (e.g. agent, theme)
  - Feature value pairs e.g.
    (TENSE PRESENT) (PERFECTIVE +)

Semantic Roles

- Verb argument semantic roles (e.g. agent, theme)
- Prepositional phrase semantic roles (e.g. instrument_with, instrument_of)
- Adjective semantic roles (e.g. attribute)
- Adverb semantic roles (e.g. manner)
- Subclause semantic roles (e.g. event)
- Numeric semantic roles (e.g. identifier “cylinder 1”)

Interlingua Features

- Tense (past, present, future)
- Modality: Ability (+,-) Possibility (+,-) Necessity (+,-), etc.
- Mood (declarative, imperative)
- Determiner Reference (Definite, Indefinite)
  Number (Singular, Plural, Mass)
  Distance (Near, Far)
- Pronoun Person (First, Second, Third)
  Number (Singular, Plural)
  Gender (Female, Male, Neuter)
- Acronym (+,-)
- Abbreviation (+,-)

KANT: Interlingua Example

(*A-LIFT (PUNCTUATION PERIOD) (TENSE PRESENT) (MOOD IMPERATIVE) (ARGUMENT-CLASS AGENT+THEME) (Q-MODIFIER (Q-MEANS_WITH (ROLE MEANS (CASE (K-WITH)) (OBJECT (O-ADJUSTABLE-LIFTING-BEAM (REFERENCE INDEFINITE) (NUMBER OR SINGULAR MASS)))))) (THEME (O-ENGINE (NUMBER SINGULAR) (REFERENCE DEFINITE))})

Interpretation Rules

- Mapping from F-structure to Interlingua:
  - (change, add, or delete information)
  - Types of mapping:
    - Head Mapping (SEM *A-LIFT) -> *A-LIFT
    - Feature promotion (PUNCTUATION PERIOD) -> (PUNCTUATION PERIOD)
    - Feature-to-feature mappings (MOOD IMP) -> (MOOD IMPERATIVE)
      (SEM *DET-AN) -> (REFERENCE INDEFINITE)
    - Feature Assignment (TENSE PRESENT) (ARGUMENT-CLASS AGENT+THEME)
    - Structural mapping OBJ -> THEME
    - PP -> Q-MODIFIER
    - Feature-to-concept mapping (SEMSLOT MEANS_WITH) -> *Q-MEANS_WITH

Interpretive Mapping

Relationship between predicate conceptual structures and syntactic structures.
- Argument Mapping: Mapping between grammatical functions and semantic roles (e.g. subject <-> agent)
- Lexical Mapping: Mapping between words and domain concepts (e.g. naguru <-> *HIT)
Alternations
Systematic semantic-syntactic correspondences
Mapping regularities within the same class of verbs
Change of state verbs
1. subject <-> agent
   object <-> theme
   a. John broke the vase.
   b. John cracked the vase.
   c. John melted the ice.
2. subject <-> theme
   a. The vase broke.
   b. The vase cracked.
   c. The ice melted.

Alternations (cont.)
Verbs of cutting:
1 a. John cut the meat.
   b. John cut at the meat.
   c. *The meat cut.
2 a. John slashed the wood.
   b. John slashed at the wood.
   c. *The wood slashed.

Change of state verbs:
3 a. *John broke at the vase.
   b. *John cracked at the vase.

Japanese Case Alternations
1. wo/de alternation: Motion-path-verbs
   a. Taro ga kawa wo oyoida
   "Taro swam down the river"
   b. Taro ga kawa de oyoida
   "Taro swam in the river"
2. kara/wo alternation: Departure-verbs
   a. Kinou uchi kara/wo denakatta.
   "(I) didn't leave the house yesterday.'
   b. Basu kara/wo oritekudasai.
   'Please get off the bus.'

Design Decisions in Domain Model Development
• Represent semantic information about objects/actions/relations in the domain
• How much domain modeling is needed for an MT system?
• How much domain modeling is feasible?
• How much domain knowledge can be extracted from existing corpora, on-line dictionaries, thesauri, etc.?
• KANT: DM is used to disambiguate sentences.
• Meaning representations (interlinguas) are instantiations of (generic) domain model concepts.

Examples
• What Domain Model frames do you need a priori?
• What do the instantiated frames look like for each example?
  Example: English to German
  – Drain the tank. "abpumpen"
  Drain the cylinder.
  – Drain the oil. "ablassen"
  Drain the fuel.

Examples (Cont.)
(*) A-DRAIN-CONTAINER
   (isa "ACTION")
   (theme "O-CONTAINER")
(*) A-DRAIN-LIQUID
   (isa "ACTION")
   (theme "O-LIQUID-SUBSTANCE")
(*) O-CONTAINER
   (isa "O-CONTAINER")
(*) O-TANK
   (isa "O-CONTAINER")
(*) O-CYLINDER
   (isa "O-CONTAINER")
(*) O-OIL
   (isa "O-LIQUID-SUBSTANCE")
(*) O-FUEL
   (isa "O-LIQUID-SUBSTANCE")
Interlingua

Drain the tank.
(*A-DRAIN-CONTAINER
(MOOD IMPERATIVE)
(theme *O-TANK))

Drain the oil.
(*A-DRAIN-LIQUID
(MOOD IMPERATIVE)
(theme *O-OIL))

Conflation: Talmy (1985)

Manner or cause is incorporated into motion verbs
– motion + manner
  The pencil rolled off the table =
  (The pencil moved off the table, rolling)
– motion + cause
  The pencil blew off the table =
  (The pencil moved off the table, from [the wind] blowing on it.)

How do we represent conflation in Domain Model?

• motion + manner concepts separately?
  – move rolling
  – roll as one concept?
• motion + cause concepts separately?
  – move blowing
  – blow as one concept?

Japanese Examples

Mitamura (1989)

action + manner:
  English -> Japanese
  (conflation) (compound verb)
  rip -> hiki-saku
        pull-tear
  chop -> kiri-kizamu
        cut mince

Japanese Examples (2)

action + manner:
  English -> Japanese
  (conflation) (no conflation)
  tap -> karuku tataku (adv + v)
        lightly hit
  slash -> satto kiru (adv + v)
        quickly cut

Japanese Examples (3)

action + instrument:
  English -> Japanese
  (conflation) (no conflation)
  punch -> genkotsu de tataku
          fist with hit
  tape -> teepu de tomeru
         tape with fix
Chinese Examples (1)
Frank Lin

• Verb-Resultant Construction in Chinese Example:

Chinese: 我 絆倒 了 他
Gloss: I trip-fall (past) him
Translation: I tripped him

Chinese Examples (2)

• Example:

Chinese: 我 打死 他
Gloss: I beat-death him
Translation: I beat him to death

Chinese Examples (3)

• Example:

Chinese: 我 喝完 了 牛奶
Gloss: I drink-finish (past) milk
Translation: I finished drinking milk

Chinese Examples (4)

• Example:

Chinese: 我 喝 光 了 牛奶
Gloss: I drink-empty (past) milk
Translation: I drank all the milk

Sinhala Examples
Sanjika Hewavitharana

1. We walked across the road.
   api avidagena paara harahaa giyyaa.
   (We went across the road by walking)
2. He drove to the town.
   ohu motar riyen nagarayata giyyaa.
   (He went to the town by a car)
3. They flew to New York.
   owun guwan vaanan New York vetha giyyaa.
   (They went to New York by plane)

One to Many Lexical Translations

• rice in Korean, Cantonese, Tamil, Sinhala, Japanese
• Family relationships in Chinese, Farsi, Telugu, Kannada
• How do we represent in the Domain Model?
  – *cooked-rice, *uncooked-rice, *rice-plant,
  – *uncooked-rice-(un)processed (Shinhala)
• What are the problems with *rice concept?
Questions?