



Interlingua MT

11-731 Machine Translation
February 7, 2007

Teruko Mitamura
Language Technologies Institute
Carnegie Mellon University



11-731: Machine Translation

1

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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)



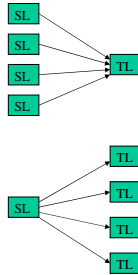
11-731: Machine Translation

2

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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



11-731: Machine Translation

3

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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



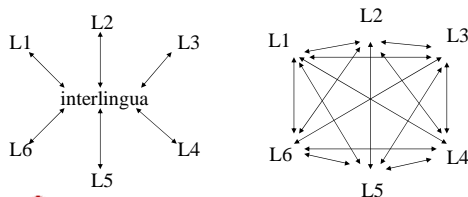
11-731: Machine Translation

4

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

The Interlingua KBMT approach

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



11-731: Machine Translation

5

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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**



11-731: Machine Translation

6

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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.



11-731: Machine Translation

7

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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



11-731: Machine Translation

8

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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.*”



11-731: Machine Translation

9

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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)



11-731: Machine Translation

10

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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



11-731: Machine Translation

11

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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



11-731: Machine Translation

12

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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)



11-731: Machine Translation

13

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

KBMT/Interlingua Examples

- Semsyn: The Doctor-Patient Domain (1986)
- KBMT-89: The Personal Computer Domain (1989)
- 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)



11-731: Machine Translation

14

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

Lexical Conceptual Structure (LCS)

- UNITRAN system (Dorr 1993)
 - English, German, Spanish
 - LCS (Jackendoff 1983, 1990) is the basis for interlingua representation
- Jackendoff, Ray S. (1983) *Semantics and Cognition*, MIT Press, Cambridge, MA
- Jackendoff, Ray S. (1990) *Semantic Structures*, MIT Press, Cambridge, MA



11-731: Machine Translation

15

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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, Perceptual, Intentional, Instrumental



11-731: Machine Translation

16

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

LCS Example

Mary died.

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



11-731: Machine Translation

17

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

LCS Example (2)

John killed Mary.

```
[Event CAUSE
  ([Thing John],
  [Events GOIdent
    ([Thing Mary],
    [Position TOWARDIdent
      ([Position ATIdent ([Thing Mary],
                          [Property DEAD])])])])])]
```



11-731: Machine Translation

18

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

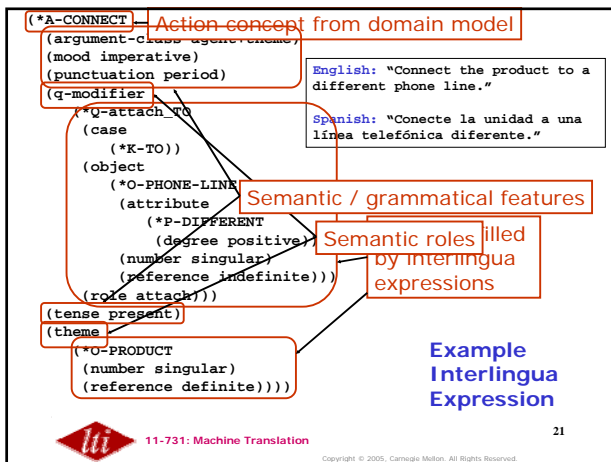
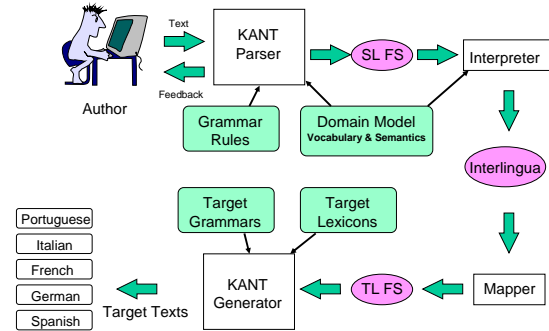
KANTOO Interlingua

- 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.

<http://www-2.cs.cmu.edu/~teruko/KANT-ir-description.html>



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"



11-731: Machine Translation

25

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

Word Sense Disambiguation (2)

- Domain Model Triples:
Turn the truck to the right.
(*A-TURN-1 (ORIENTED_TO *O-RIGHT-1))

The deposits turn into sludge.
(*A-TURN-2 (RESULT_INTO *O-SLUDGE))



11-731: Machine Translation

26

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

Word Sense Disambiguation (3)

INTERLINGUA for *A-TURN-1:

Turn the truck to the right.

(*A-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))))))



11-731: Machine Translation

27

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

Word Sense Disambiguation (4)

INTERLINGUA for *A-TURN-2:

The deposits turn into sludge.

(*A-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))))))



11-731: Machine Translation

28

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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



11-731: Machine Translation

29

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

KANTOO IL design

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



11-731: Machine Translation

30

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

Head Concepts

Concept Prefix	Denotation
*A	Action (verbal) e.g. *A-MAKE-SURE
*O	Object (nominal) e.g. *O-FAX-MACHINE
*M	Manner (adverbial) e.g. *M-THOROUGHLY
*P	Property (adjectival) e.g. *P-MULTI-COLORED
*K	(preposition) *K-ABOVE
*INT	Intensifier e.g. *INT-VERY
*CONJ	Conjunction e.g. *CONJ-THAT
*QUANT	Quantifier e.g. *QUANT-SOME
*PROP	Proper (proper nouns) e.g. *PROP-VERDANA
*SYM	Symbol (e.g. typographic symbols) e.g. *SYM-EXCLAMATION-POINT
*U	Unit (measurement) e.g. *U-AMPERE
*C	Crystal (complex structure, untagged) *C-DECIMAL-NUMBER
*G	Grammatical e.g. *G-GAPPED-ARGUMENT
*Q	sem-role_preposition e.g. Q-goal_INT0
*S	Structured (complex structure, tagged)
*SP	Special (domain-specific phrase) e.g. *SP-NOT-AVAILABLE



11-731: Machine Translation

31

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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



11-731: Machine Translation

32

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

“The manual was printed.”

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



11-731: Machine Translation

33

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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



11-731: Machine Translation

34

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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
(*O-PRINTER))
(number singular)
(person third)
(reference definite)))



11-731: Machine Translation

35

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

Use the printer, if it is clean.

(*A-USE
(argument-class agent+patient) (mood imperative)
(patient
(*O-PRINTER
(number singular) (reference definite)))
(punctuation period)
(qualifier
(*G-QUALIFYING-EVENT
(event
(*A-BE-PREDICATE
(attribute
(*P-CLEAN
(degree positive)))
(mood declarative) (predicate-role attribute)
(tense present)
(theme
(*G-PRONOUN
(antecedent
(*O-PRINTER))
(number singular)
(person third)
(reference definite)))
(extent
(*CONJ-if)))
(tense present))



11-731: Machine Translation

36

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

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_INT0
(case
(*K-INT0))
(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.aitcnet.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?

