


Overview of KANTOO MT (Analysis and Generation)

11-731 Machine Translation




11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

1

Source Analysis




11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

2

KANT: System Architecture

- Modules
 - segmenter, morphology, parser, interpreter, filter
- Knowledge
 - tag database (DTD), lexicon, grammar, semantic model
- Levels of Representation
 - input string, syntactic f-structure, interlingua, ...



11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.


3

KANT Analyzer Data Flow

```

    graph LR
      A[SGML text] --> B[segmenter]
      B --> C[Sentence]
      C --> D[morphology]
      D --> E[Stems+Affixes]
      E --> F[parser]
      F --> G[Set of F-Structures]
      G --> H[filter]
      H --> I[Single F-Structure]
      I --> J[interpreter]
      J --> K[Interlingua]
    
```

In Software Engineering, this kind of data flow is referred to as a *transform flow* or *transform mapping*




11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

4

| | |
|---|------------------------------------|
| <p>Analyzer</p> <ul style="list-style-type: none"> – Tokenizer (String -> Tokens) – Lexifier (Tokens -> Frames/FS) <ul style="list-style-type: none"> Morphonizer (String -> Morpheme) Wordifier (String -> Frame/FS) Idiomifier (Tokens+Head -> Frame/FS) Morphsemizer (Description -> Def) – Syntaxifier (Frames/FS -> Frame/FS) <ul style="list-style-type: none"> Parser (Frames/FS -> Frame/FS) – Disambiguator (Frame/FS -> Frame/FS) <ul style="list-style-type: none"> Domo_Search (Domokey -> PP_Semrole) – Interpreter (Frame/FR -> Frame/IR) | <p>Details: Modules</p> |
|---|------------------------------------|




11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

5

| | | |
|--|---|--|
| <p>Database</p> <ul style="list-style-type: none"> DMK Morphonetic Morphsemantic Grammar LR_State_Table Domo Ambig_Heuristic Interlingua_DB | <p>Module</p> <ul style="list-style-type: none"> Lexifier Morphonizer Morphsemizer Parser Parser Domo_Search Disambiguator Interpreter | <p>Details: Knowledge Bases</p> |
|--|---|--|




11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

6

| <u>Knowledge</u> | <u>Type</u> | <u>Database</u> | Details: Knowledge Types |
|----------------------|-------------|-----------------|---|
| Idiom_Dictionary | Map | DMK | |
| Irregular_Dictionary | Map | DMK | |
| Word_Dictionary | Map | DMK | |
| Morphonetic | Rules | Morphonetic | |
| Morphsemantic | Rules | Morphsemantic | |
| Domokeys | Map | Domo | |
| Generalizations | Map | Domo | |
| Grammar | Rules | Grammar | |
| LR_State_Table | Map | LR_State_Table | |
| Ambiguity | Rules | Ambig_Heuristic | |
| Interpreter | Rules | Interlingua_DB | |
| Interlingua_Tree | Rules | Interlingua_DB | |

Map: accesses stored knowledge associated with a key (e.g. term)
Rules: transform a structure; usually PATRICK or C code




11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

7

System Architecture

- System Characteristics
 - Interactive grammar checking (author's workstation)
 - Batch translation on server machines
 - Regular knowledge updates




11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

8

Segmentation

- From SGML text to individual input sentences
- Separate "debris" from translatable text
- Break paragraphs into sentences
- Behavior controlled by separate tag data file & list of allowable abbreviations




11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

9

Preprocessing & Morphology

- Canonicalization (e.g., case)
- Chunking (e.g., idioms, technical phrases)
- Identify possible root forms plus affixes (morphemes)
- Lexical lookup (filter out impossible root forms)
- Create/annotate lexical structure




11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

10

Syntactic Parsing

- What types of knowledge?
- What levels of representation?
- Theory-driven vs. domain-driven systems
- KANT: non-deterministic, exhaustive LR parsing (Tomita)
- For each input, possibly multiple f-structures




11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

11

Semantic Restrictions

- What constructions are ambiguous? What knowledge can be used to disambiguate?
- Top-down (rational) model
 - Every concept in the domain is fit into a complete, comprehensive domain model (hierarchy)
 - Extremely precise and detailed
 - Extremely expensive, too (CYC)



11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

12

Semantic Restrictions [2]

- Bottom-up (empirical) model
 - Include only concepts & relations that are necessary to disambiguate
 - Create parent classes only to support generalization where needed
 - Less precise, more approximate
 - Much less expensive
 - Can give better cost benefit



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

13

Semantic Interpretation

- Integration with syntactic processing?
- How domain/language independent?
- KANT: Mapping f-structures to interlingua



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

14

Filtering Algorithms

- Automatic methods
 - General preference rules
 - Domain-specific preference rules
 - Phrasal vs. compositional parses
"Oil flows through the bearing seal."
- Interactive methods
 - Author disambiguation
 - Input annotation
- KANT: automatic & interactive



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

15

More Details

- See the papers on the KANT web:
<http://www.lti.cs.cmu.edu/Research/Kant/>
 - System Architecture: 1-4
 - Corpus Analysis: 5, 11, 14
 - Disambiguation: 7, 12, 24
 - Evaluation: 8, 17, 20, 23
 - Controlled Input: 13, 15, 21, 25



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

16

Target Generation



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

17

NLG Ingredients

- A representation of the input (probably not human-friendly)
- Knowledge of the domain
- Knowledge of the target language
- A human-friendly output format:
 - documents, reports, explanations, help messages, technical instructions, etc.



Carnegie Mellon
School of Computer Science


11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

18

6 Basic NLG Tasks

1. Content Determination
what information should be conveyed?
2. Discourse Planning
order & structure of message set
3. Sentence Aggregation
grouping messages into sentences
4. Lexicalization
words & phrases for concepts, relations
5. Referring Expression Generation
words & phrases for entities
6. Linguistic Realisation
syntax, morphology, orthography

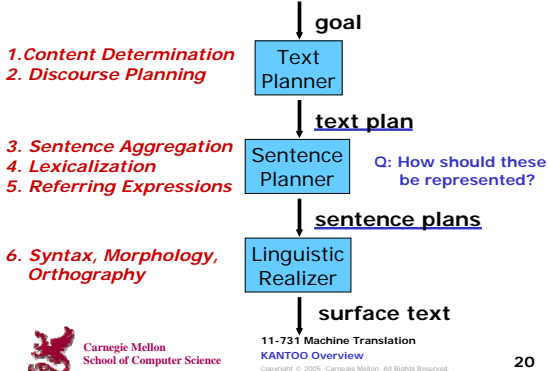


Carnegie Mellon School of Computer Science

11-731 Machine Translation
KANTOO Overview
Copyright © 2005, Carnegie Mellon. All Rights Reserved.

19

Typical 3-Module Architecture




1. Content Determination
2. Discourse Planning

3. Sentence Aggregation
4. Lexicalization
5. Referring Expressions

6. Syntax, Morphology, Orthography

Q: How should these be represented?




Carnegie Mellon School of Computer Science

11-731 Machine Translation
KANTOO Overview
Copyright © 2005, Carnegie Mellon. All Rights Reserved.

20

Text Plans

- Common representation : tree
 - Leaf nodes = messages
 - Internal nodes = message groupings
- Simple text plans: templates OK
- Complex text plans: require full representation language (e.g., TAMERLAN, DIOGENES)




Carnegie Mellon School of Computer Science

11-731 Machine Translation
KANTOO Overview
Copyright © 2005, Carnegie Mellon. All Rights Reserved.

21

Sentence Plans

- Simple: templates (select & fill)
- Complex: abstract representation (SPL: Sentence Planning Language)



Carnegie Mellon School of Computer Science


11-731 Machine Translation
KANTOO Overview
Copyright © 2005, Carnegie Mellon. All Rights Reserved.

22

Example SPL Expression

```
(S1/exist
:object (01/train
:cardinality 20
:relations ((R1/period
:value daily)
(R2/source
:value Aberdeen)
(R3/destination
:value Glasgow))))
```

There will be 20 trains to Glasgow




Carnegie Mellon School of Computer Science

11-731 Machine Translation
KANTOO Overview
Copyright © 2005, Carnegie Mellon. All Rights Reserved.

23

Content Determination

- Messages (raw content)
- User Model (influences content)
- Is Reasoning Required?
Find a train from Aberdeen to Leeds
(It requires two trains to get there)
- Deep Reasoning Systems
 - represent the user's goals as well as any immediate query
 - utilize plan recognition & reasoning



Carnegie Mellon School of Computer Science

11-731 Machine Translation
KANTOO Overview
Copyright © 2005, Carnegie Mellon. All Rights Reserved.

24

Discourse Planning

- Structure messages into a coherent text
- Example: start with a summary, then give details
- Discourse relations, e.g.:
 - elaboration: *More specifically, X*
 - exemplification: *For example, X*
 - contrast / exception: *However, X*
- Rhetorical Structure Theory (RST)



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

25

Sentence Aggregation

- No aggregation (1 sentence / message)
- Relative Clause
..which leaves at 10am
- Conjunction
..and the next train is the express
- Combinations
*..and the next train is the express
which leaves at 10am*



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

26

Lexicalization

- Choosing words to realize concepts or relations
- Example:
(action/change
(measure outside_temperature)
(delta (quantity/deg_F -10)))

The temperature dropped 10 degrees



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

27

Lexical Selection Rules

```
(*A-INGEST
 (AGENT *O-BOB)
 (PATIENT *O-MILK)) => "drink"

(*A-INGEST
 (AGENT *O-BOB)
 (PATIENT *O-CHOCOLATE)) => "eat"
```



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

28

Case Creation

- Additional structure is required to realize the meaning of the semantic representation

```
(*A-KICK
 (AGENT *O-JOHN)
 (PATIENT *O-BALL))
```

"John propelled the ball with his foot"



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

29

Case Absorption

- Word chosen to realize a semantic head also implies the meaning conveyed by a semantic role

```
(*A-FILE-LEGAL-ACTION
 (AGENT *O-BOB)
 (PATIENT *O-SUIT)
 (RECIPIENT *O-ACME))
```

"Bob sued Acme"



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

30

Referring Expression Generation

- Initial introduction
A man in the park looked up
- Pronouns
He saw a bird fly over
- Definite Descriptions
The man covered his head with a newspaper



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

31

Fixing Robot Text

- Start [the engine]_i and run [the engine]_i until [the engine]_i reaches normal operating temperature
- Start [_i] and run [the engine]_i until [it]_i reaches normal operating temperature
- Second example introduces *ellipsis* and *anaphora*



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

32

Journalistic Style

"[A dissident Spanish priest](#) was charged here today with attempting to murder the Pope. [Juan Fernandez Krohn](#), aged 32, was arrested after [a man armed with a bayonet](#) approached the Pope while he was saying prayers at Fatima on Wednesday night. According to the police, [Fernandez](#) told the investigating magistrates today, [he](#) trained for the past six months for the assault. If found guilty, [the Spaniard](#) faces a prison sentence of 15-20 years."
(Brown and Yule, 1983)



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

33

Other Readings

- KBMT-89 book chapter on generation (algorithms & control)
- MT journal paper (acquiring MT knowledge sources)
- Paper on Interlingua design
- Paper on Turkish generation
- Paper on Chinese generation



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

34

Generator Software Development

- Defining the input /domain requirements
- Defining the output / quality requirements
- Functional specification
- Selecting a software architecture
- Case Study: The KANT Generator



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

35

Input / Domain Requirements

- Grain size: sentences, paragraphs, or texts?
- Errors: graceful degradation, no output, source?
- Controlled vs. general domains
Is the output controlled?
- Definition by construction/corpus



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

36

Domain Analysis

- What source phenomena?
- Basic representational elements? (BNF for input representation)
- What combinations thereof? (hard to spec!)
- Derive input specification (e.g. IR Spec)



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

37

Existing Knowledge Sources

- Dictionaries (multi-lingual?)
- Bilingual, aligned corpora (e.g., Lonsdale's BiKwik)



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

38

Output / Accuracy Requirements

- Lexical accuracy
- Grammatical acceptability
- Domain-specific phenomena
 - hard constructions (passive, reference, aspect, etc.)
 - complex combinations
 - tagging in the input (e.g., determiner stranding)



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

39

Requirements Specification

- Lexical, Structural, Morphological processes
- What levels of processing are necessary?
 - the "eight levels" model
- Knowledge sources required?
- Load/run/update characteristics?
- Support tools/processes?



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

40

KANT: Use of Existing Knowledge

- Parts database, some multilingual entries
- Translator's glossaries / dialect issues



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

41

Output Quality: What's good enough?

- No loss of meaning
- Little or no post-editing required
- Cost of MT + postediting should be less than human translation



Carnegie Mellon
School of Computer Science


11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

42

KANT: Generator Architecture (1)

- Modules
 - Mapper, GenKit, Morphe, CODA
- Top-down, recursive algorithm
- Knowledge
 - lexical rules, structural rules, grammar, morph classes/rules, CODA rules




11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

43

KANT: Generator Architecture (2)

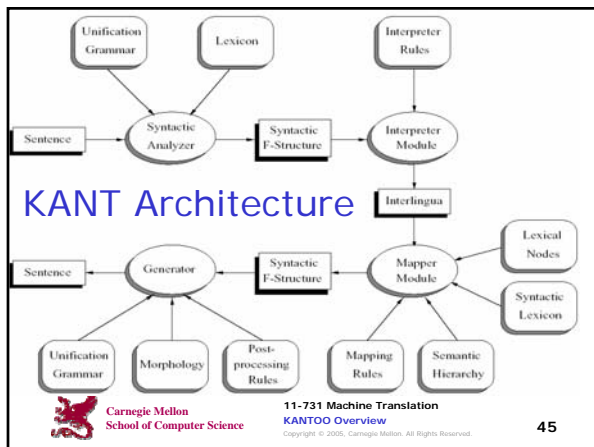
- Levels of Representation
 - input string, syntactic f-structure, interlingua
- System Characteristics
 - non-interactive
 - batch translation on server machines
 - regular knowledge updates (sync with CTE)



11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.


44



Example Interlingua

*** Mapper: Input:

```
(*O-SCANNER
(attribute
(*P-EXISTING
(degree positive)))
(number singular)
(reference definite)
(standalone-phrase +))
```



11-731 Machine Translation
KANTOO Overview


Copyright © 2005, Carnegie Mellon. All Rights Reserved.

46

Example F-Structure

*** Mapper: Output: (

```
(agr ((gender masc) (number sg) (person 3)))
(cat n)
(det ((cat det) (root el) (type def)))
(non-phra-modifier ((cat adj) (root existente)))
(root escáner)
(type reg))
```




11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

47

Linearized F-Structure

```
(((root "@cap"),
(agr ((gender masc) (number sg)
(person 3)))
(cat det)
(gender masc)
(number sg)
(root el)
(type def)),
(agr ((gender masc) (number sg)
(person 3)))
(cat adj)
(change si)
(gender masc)
(number sg)
(root existente)
(type reg))]
```



11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

48

Morphology & Post-Proc

*** Morphoee: Input: (
(agr ((gender masc) (number pl) (person 3)))
(cat adj)
(change si)
(gender masc)
(number pl)
(root existente)
(type reg))
*** Morphoee: Output: existentes
*** Codafier: Output: Los escáneres existentes



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

49

Demo



Carnegie Mellon
School of Computer Science

11-731 Machine Translation
KANTOO Overview

Copyright © 2005, Carnegie Mellon. All Rights Reserved.

50