Practical Knowledge-Based Machine Translation: Experience with the KANT System

Teruko Mitamura
Eric H. Nyberg, 3rd
Jaime Carbonell

Center for Machine Translation
Carnegie Mellon University
Taxonomy of MT Tasks

"General" Translation

Translation for Assimilation

Scanning for Relevance

Extracting Data

Indexing and Store

Translation for Dissemination

Informal & Interaction (e.g. Speech-Speech MT)

Event Analysis (e.g. trend detection)

High Volume per Domain

Publication Quality

One of a Kind Document

Center for Machine Translation
Assimilation vs. Dissemination

ASSIMILATION

Source Languages -> Target Language

* Any Language
* Any Style
* Almost Any Topic
* General-Purpose MT
* No Semantic Analysis
* Requires Post-Editing

DISSEMINATION

Source Language -> Target Languages

* One Source Language
* Controlled Style
* Single Topic or Domain
* Special-Purpose MT
* Full Semantic Analysis
* No Post-Editing
KANT Design Objectives

- No Post-Editing of Translation Output

- Minimization of Human Help in Disambiguation (Intervention-Free Translation)

- Cost-Effective Multiple-Language Output

- Efficient On-Line Translation

- Efficient System Development

- Cost-Effective System Maintenance and Extension
KANT Design Solutions

• Controlled input language allows intervention-free translation.

• Knowledge-based parsing, interpretation, and generation imply no hard-wired limitations in coverage.

• Powerful rule formalisms for analysis and generation provide effective application of linguistic knowledge.

• Knowledge pre-compilation turns linguistic knowledge into efficient run-time data structures.

• Semi-automated tools for knowledge acquisition keep cost of maintenance and extension low.
Interlingua Translation

Source Text → Analysis → Interlingua → Generation → Target Text
KANT Architecture

Source Language Analysis

CTE Text → PARSER → F-structure → INTERPRETER

TARGET LANGUAGE GENERATION

French Text → GENERATOR → F-structure → MAPPER

Interlingua
KANT Runtime Knowledge

PARSER

INTERPRETER

Source Text

Phrasal Lexicons

Run-Time Grammar

Syntactic Lexicon

Run-Time Semantic Restrictions

Mapping Rules

F-structure

Interlingua

GENERATOR

MAPPER

Target Text

General Mapping Rules

Run-Time Target Lexicon

Run-Time Morphology

Run-Time Grammar

Post-Proc Rules

F-structure
The KANT Architecture

Source Text

Analysis Module

Parser

Source Grammar
Source Lexicon

Source F-Structure

Interpreter

Source Mapping Rules

Domain Model

Interlingua

Mapper

Domain Model

Mapper

Target Mapping Rules
Target Lexicon
Target Grammar
Target F-Structure
Target Text

Generation Module

Generator

Target Text1

Target Text2

Center for Machine Translation
Safety Warnings

Read the “General Installation Information” section of this manual. Then, follow the instructions in the “Safety Warnings” section.

In order to prevent a fire hazard, do not overload AC outlets.

In the following cases, TV sets can overheat:

1. The ventilation slots are blocked.
2. The TV set is placed in a built-in enclosure.

Periodically clean the ventilation slots with your vacuum cleaner.

If the TV set has been dropped, a shock hazard may exist. In this case, unplug the TV set. Then call your dealer.
Conseils de sécurité

Consulter la section de ce manuel intitulée “Renseignements pour installation”. Ensuite, se conformer aux instructions figurant à la section intitulée “Conseils de sécurité”.

Afin d’éviter tout risque d’incendie, ne jamais surcharger les prises CA.

Dans les cas suivants, un téléviseur peut surchauffer:

1. La grille de ventilation est bloquée.
2. Le téléviseur est placé dans un coin renfoncé.

Dépoussiérer périodiquement la grille de ventilation à l’aide d’un aspirateur.

La chute du téléviseur peut provoquer un risque de choc électrique. En ce cas, débrancher le téléviseur. Ensuite faire appel au détaillant.
Avvertenze di sicurezza

Consultare la sezione “Informazioni per l’installazione” di questo manuale. Poi, conformarsi alle istruzioni nella sezione intitolata “Avvertenze di sicurezza”.

Per evitare un’eventuale emergenza di fuoco, non sopraccaricare le prese CA.

Nei casi seguenti, i televisori possono soprarriscaldarsi:

1. La griglia di ventilazione è blocata.

2. Il televisore è posto in un angolo rinforzato.

Periodicamente, togliere la polvere alla griglia di ventilazione con l’aiuto del vostro aspirapolvere.

La caduta del televisore può provocare un rischio d’urto elettrico. In questo caso, disinserire il televisore. Poi, fare appello al vostro dettagliante.
Sicherheitsbestimmungen


Fernsehgeräte können in den folgenden Fällen überhitzen:

1. Die Kühlschlitze sind blockiert.
2. Das Fernsehgerät steht in einem Einbauschrank.

Reinigen Sie regelmäßig die Kühlschlitze mit dem Staubsauger.

Wenn Sie das Fernsehgerät fallenlassen, kann die Gefahr eines Elektroschocks bestehen. Ziehen Sie in diesem Fall den Netzstecker. Verständigen Sie dann Ihren Kundendienst.
「安全の注意」

このマニュアルの「一般設置情報」の章を読んで下さい。それから「安全の注意」の章の指示に従って下さい。

火災の危険を防ぐために、コンセントに電流の負担をかけすぎないで下さい。
次の様な場合はテレビがオーバーヒートすることがあります。

1. 通気孔がふさいでいる。
2. 作りつけの畳まれた場所にテレビがある。

定期的に掃除機で通気孔を掃除して下さい。

テレビを落とすと、電気ショックの危険があるかもしれません。その場合はテレビのプラグを抜いて下さい。それから販売店に連絡して下さい。
CMU’s MT Factory

New Domain

New Language

Domain Expert

Corpora

Ontos

Corpus Analysis Tools

Lexicon

CLFG Grammar

Linguist

Dictionary Compiler

Generation Compiler

Parsing Compiler

Development

Runtime

New Data for each MT system

I-rules

M-rules

Lexicon

Generation Grammar

Parsing Grammar

Fully Reusable Runtime Modules

Universal Parser

Source Language Analysis

Interpreter

Source Language Interpretation

Mapper

Target Language Semantic ➔ Syntax

Generator

Target Language Syntax ➔ Text

Morpheme

Morphological Analysis/Generation

Coda

Punctuation, Accents, Euphonics, ...

= KBMT
Intermediate Data Structures (1)

Source Sentence:

"In the following cases, TV sets may overheat."

Source F-structure:

((MOOD DEC) (GAP -) (FORM ROOTFORM) (ROOT "overheat")
 (CAT V) (VALENCY INTRANS)
 (PRE-MOD-PP
   ((ROOT "in") (CAT P)
    (OBJ
     ((SEM *O-CASE) (NUMBER PL) (ROOT "case") (CAT N)
      (COUNT +)
      (DET
       ((ROOT "the") (CAT DET))))
     (ADJ-MOD
      ((ROOT "following") (CAT ADJ)
       (SEMSLOT ATTRIBUTE)))))))
 (SUBJ
  ((SEM *O-TELEVISION) (NUMBER PL) (ROOT "set") (CAT N)
   (COUNT +)))
 (MODAL
  ((ROOT "may") (CAT ANOM.FIN.))))
Intermediate Data Structures (2)

Source Sentence:
"In the following cases, TV sets may overheat."

Interlingua Text:

(*E-OVERHEAT
  (MOOD DEC)
  (MODAL POSSIBILITY)
  (CONDITION
    (*O-CASE
      (NUMBER PL)
      (REFERENCE DEFINITE)
      (ATTRIBUTE (*P-FOLLOWING))))
  (THEME
    (*O-TELEVISION
      (NUMBER PL))))
Intermediate Data Structures (3)

French Target F-structure:

```
((TENSE PRESENT) (VOICE ACTIVE) (MOOD INDICATIVE)
 (MODAL +) (CAT VERB) (ROOT "pouvoir")
 (SUBJ
  ((DET-FREE +) (NUMBER-IMMUNITY +) (CAT NOUN)
   (ROOT "te’le’viseur") (DET ((ROOT UN)))
   (AGR
    ((PERSON 3) (NUMBER SG) (GENDER M))))))
 (S-MODIFIER ((PPADJUNCT
   ((PREP ((ROOT "dans")))
    (P-OBJ
     ((CAT NOUN) (ROOT "cas")
      (DET ((ROOT LE)))
      (MODIFIER
       ((CAT ADJ) (ROOT "suivant")))
      (AGR
       ((PERSON 3) (NUMBER PL) (GENDER M))))))))
 (XCOMP
  ((ROOT "surchauffer") (CAT VERB))))
```

French Target Raw String:

"@CAP dans LES cas suivantS COMMA UN téléviseur peut surchauffer PERIOD"

French Target Sentence:

"Dans les cas suivants, un téléviseur peut surchauffer."
KANT Example Translation

“Operate the engine until the engine reaches normal operating temperature.”

(*E–OPERATE
   (MOOD IMP)
   (TERMINATION–EVENT
    (*E–REACH
     (MOOD DEC)
     (THEME
      (*O–ENGINE
       (NUMBER SG)
       (REFERENCE DEFINITE)))
    (GOAL
     (*O–TEMPERATURE
      (BELONGS–TO (*O–OPERATING))
      (ATTRIBUTE (*P–NORMAL)))))
   (THEME
    (*O–ENGINE
     (REFERENCE DEFINITE))))

“Faire tourner le moteur jusqu’à ce que le moteur atteigne la température de marche normale.”
KANT Example Translation

“When the temperature reaches the cloud point, a wax forms in the fuel.”

(*E-FORM
  (MOOD DEC)
  (WHEN-EVENT
    (*E-REACH
      (MOOD DEC)
      (THEME
        (*O-TEMPERATURE
          (NUMBER SG)
          (REFERENCE DEFINITE)))
      (GOAL
        (*O-CLOUD-POINT
          (REFERENCE DEFINITE))))
    (THEME
      (*O-WAX
        (NUMBER SG)
        (REFERENCE INDEFINITE)))
    (LOCATION
      (*O-FUEL (REFERENCE DEFINITE))))

“Lorsque la température baisse jusqu’au point de trouble, une paraffine se forme dans le carburant.”
Building a KANT Application

1. Corpus Preparation
2. Corpus Analysis
3. Source Vocabulary Creation
4. Source Grammar Creation
5. Domain Model Construction
6. Vocabulary Translation
7. Target Grammar Creation
8. Integration and Testing
9. Knowledge Refinement
Corpus Preparation

• Getting the right customer corpus sample
  – Across all products
  – Across all document types
  – With as much corresponding target language
    text as possible

• Deformatting the files
  – Neutralize/canonicalize different character
    representations, markup codes
  – Break into sentences

• Create corpus KWIC browsers
  – For the source language
  – For the target language(s)
Example KWIC Browser Output

Term to show? ripper tip

iter valve ( 21 ) is used in the ripper tip control valve to control ( limited pressure limiter valve for the ripper tip circuit ) . When the ripper tip control valve is in an operational position, pump oil through the ripper tip valve spool ( 21 ) . The ripper tip control valve goes to inlet passages .

( 20 ) makeup valve . ( 21 ) ripper tip valve spool . ( 22 ) ripper tip control valve . ( 23 ) load chamber . ( 24 ) shuttle valve . ( 25 ) ripper tip valve spool ( 21 ) and out of the ripper tip circuit .

This return oil comes into the ripper tip control valve and goes around the common tank passages in the ripper tip control valve ( 22 ) to the ripper tip control valve spool ( 21 ) and out of the ripper tip control valve .

The return oil comes into the ripper tip control valve and goes around the common tank passages in the ripper tip control valve , ripper lift control valve ( 22 ) to control control valve ( 22 ) to control the hydraulic cylinder rods for movement .

Drift tests for ripper tip cylinders:

1. Tip the machine off the engine and watch the ripper tip cylinder rods for movement .
2. Tip the machine off the engine and watch the ripper tip cylinder rods for movement .
3. Install floor plate .
4. Pressure limiter valve for the ripper tip circuit is 22 /th/ 750 /pm/ 86 psi .
5. Remove plug ( 3 ) from the ripper tip control valve .
6. Turn adjuster .

Tools needed: 6v4161 circuit

Equipment .
Corpus Analysis

- Single-Word Terms
  - Compile a list of all words with number of occurrences
  - Canonicalize via morphological analysis

- Phrasal Terms
  - Define heuristics for locating phrases
  - Compile a list of all candidate phrases with number of occurrence
  - Canonicalize via morphological analysis
Single Word Terms

630186 the
185352 and
178946 to
140407 of
127776 in
103206 is
95967 a
69145 valve
68193 for
65787 oil
63046 on
56781 engine
51778 or
46661 from
44430 be
44259 pressure
43906 with
41669 remove
38728 install
38111 control
34131 fuel
33405 pump
Phrasal Terms

6092 lb ft
4889 cooling system
3966 fuel injection
3862 parking brake
3410 relief valve
2926 cylinder head
2924 test force
2886 oil pressure
2789 control valve
2634 torque converter
2587 service hours
2491 o-ring seal
2421 hydraulic oil
2373 caterpillar dealer
2307 hydraulic tank
2306 personal injury
2230 engine speed
2211 outside diameter
2188 cylinder block
2089 engine oil
2037 lift truck
1945 low idle
Source Vocabulary Creation

• Automatic Processing
  – Tagging word occurrences with possible parts of speech
  – Automatic creation of a lexicon from a word list and default templates

• Human Processing
  – For each general word entry, KWIC browse corpus to verify POS, other syntactic features; add examples
  – Add missing terms, classes of terms (completeness)
  – Check for ambiguity (coherence)

• Automatic Processing, Take 2
  – Refine phrasal vocabulary based on updated term vocabulary
  – Refine phrasal vocabulary based on frequency
Sample Lexicon Entry

((:ROOT "rip")
 (:POS V)
 (:CONCEPT *A-RIP)
 (:SYL-DOUBLE +)
 (:SYN-FEATURES (VALENCY TRANS INTRANS))
 (:NOTE
  (:SENSE
   "Technical term: to slash into with a ripper"
   "There are several ways to rip hard spots and boulders."
   "Rip downhill whenever possible."
   "Do not rip and doze at the same time."))
 (:COMMENT "not sure of definition")
 (:ACTION :HYD :TECH)
 (:FREQUENCY 106 368)
 (:UPDATED (20 29 18 26 6 1992) "catguest")
Example Phrase Refinements

(NIPPONDENSO PRODUCE) (N V)
(JOG ENGINE STARTER) (V N N)
(ANNUNCIATE OVERCRANK SHUTDOWN) (V N N)
(LINE BEHIND TRACTOR) ((V N) (PREP ADV) (N))
(ACTUATOR STILL) (N ADV)

3 (ASBESTOS-FREE FRICTION)
   (ASBESTOS-FREE FRICTION MATERIAL)

47 (ATAAC TRUCK)
   (ATAAC TRUCK ENGINE)

4 (AUTOMATIC BLADE)
   (AUTOMATIC BLADE CONTROL)

2 (BASE OIL)
   (BASE OIL VISCOSITY)
Source Grammar Creation

- Analysis of source documents (for style, orthography, mark-up, etc.)

- Creation of language standard (e.g., Controlled English)

- Refinements to source lexicon

- Creation of an LFG syntactic grammar, mapping rules
Domain Model Construction

- Bottom Up: from words to concepts
  - Mapping POS to concept types
  - Capturing concept relations
  - Building semantic restrictions

- Top Down: from the customer domain expert
  - Customer’s perspective on the domain (not always congruent with translation)
  - Understanding potential concept relations
Example Object Frame

(*O-TEMPERATURE
 (is-a *O-ABSTRACT-OBJECT)
 (belongs-to
   *O-FUEL *O-OPERATING *O-FUEL-SUPPLY
   *O-THERMOSTAT *O-ENGINE-COMPARTMENT
   *O-ENGINE *O-AREA *O-OPERATION)
 (under *O-TEMP-MEASUREMENT-UNIT)
 (attribute
   *P-AMBIENT *P-HIGH *P-NORMAL
   *P-APPROPRIATE *P-AVERAGE *P-OUTSIDE
   *P-FREEZING))
Vocabulary Translation

- Automated processing:
  - Corpus preparation (target language documents)
  - Corpus alignment (with corresp. English documents)
  - Bilingual Browsing for technical translations
  - Monolingual Browsing for general translations
  - Automatic lexicon creation

- Human processing:
  - Review of technical and general translations by customer translations expert
  - Refinement of automatically created lexicon
Target Grammar Creation

- Analyze existing target language documents for style, orthography, etc.
- Consult with customer expert translators
- Propose standard rules of target language usage (e.g., Controlled French)
- Make refinements to target vocabulary as necessary
- Construct and test LFG grammar, mapping rules
Example Target Mappings

(glex *e-reach
 (pattern (and (theme *o-temperature)
               (or (goal *o-cloud-point)
                   (goal *o-pour-point))))
 (syn "baisser")
 (map (theme subj)
      (goal (ppadjunct
             ((prep ((root jusqu’a)))
              (p-obj *)))))

(glex *e-reach
 (pattern (and (theme *o-engine)
               (goal *o-temperature))))
 (syn "atteindre")
 (map (theme subj)
      (goal obj)))
Integration and Testing

- Build, benchmark, optimize complete system
- Test on “live” text
- Test in customer’s computing environment
<table>
<thead>
<tr>
<th>Name</th>
<th>Responsibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>James Altucher</td>
<td>Analyzer Software, Domain Model</td>
</tr>
<tr>
<td>Kathryn Baker</td>
<td>Controlled Technical English</td>
</tr>
<tr>
<td>Nicholas Brownlow</td>
<td>Domain Model, Corpus Analysis Tools</td>
</tr>
<tr>
<td>Bertrand Damiba</td>
<td>French Lexicon</td>
</tr>
<tr>
<td>Henry Debusmann</td>
<td>French Lexicon</td>
</tr>
<tr>
<td>Alex Franz</td>
<td>English Grammar, Interpreter, Analysis Mapping Rules</td>
</tr>
<tr>
<td>Mildred Galarza</td>
<td>Spanish Generation</td>
</tr>
<tr>
<td>Stefan Geissler</td>
<td>German Grammar, Lexicon, and Mapping Rules</td>
</tr>
<tr>
<td>Susan Holm</td>
<td>English Lexicon</td>
</tr>
<tr>
<td>Kathi Iannamico</td>
<td>English Lexicon</td>
</tr>
<tr>
<td>Ingrid Fischer</td>
<td>German Grammar, Lexicon, Morphology and Mapping Rules</td>
</tr>
<tr>
<td>Pamela Jordan</td>
<td>Analyzer Software, English Grammar</td>
</tr>
<tr>
<td>Todd Kaufmann</td>
<td>Knowledge Acquisition Tools</td>
</tr>
<tr>
<td>Kevin Keck</td>
<td>Spanish Generation</td>
</tr>
<tr>
<td>Kristina Keenan</td>
<td>French Lexicon</td>
</tr>
<tr>
<td>Ken Lambert</td>
<td>German Generation</td>
</tr>
<tr>
<td>Sarah Law</td>
<td>French Lexicon</td>
</tr>
<tr>
<td>John Leavitt</td>
<td>Generation Software, French and Spanish Generation</td>
</tr>
<tr>
<td>Daniela Lonsdale</td>
<td>English Lexicon, German Generation</td>
</tr>
<tr>
<td>Deryle Lonsdale</td>
<td>French Grammar, Lexicon, and Mapping Rules</td>
</tr>
<tr>
<td>Sebastian Meller</td>
<td>German Generation</td>
</tr>
<tr>
<td>Jeanne Mier</td>
<td>French and Spanish Generation</td>
</tr>
<tr>
<td>Mark Miller</td>
<td>English Lexicon</td>
</tr>
<tr>
<td>Venkatesh Narayan</td>
<td>Domain Model</td>
</tr>
<tr>
<td>Michelle Vanni</td>
<td>Italian Grammar, Lexicon, Morphology and Mapping Rules</td>
</tr>
<tr>
<td>Willy Walker</td>
<td>English Lexicon, Knowledge Acquisition Tools</td>
</tr>
</tbody>
</table>