THE KANT SYSTEM: FAST, ACCURATE, HIGH-QUALITY TRANSLATION IN PRACTICAL DOMAINS

Eric H. Nyberg III
Teruko Mitamura
Center for Machine Translation
Carnegie Mellon University
Pittsburgh, PA 15213 USA

Project Note / Demonstration

Major Topic Area: 5*c (Machine Translation)
Minor Topic Area: 7*a (Language Engineering)

Abstract

Knowledge-based interlingual machine translation systems produce semantically accurate translations, but typically require massive knowledge acquisition. Ongoing research and development at the Center for Machine Translation has focussed on reducing this requirement to produce large-scale practical applications of knowledge-based MT. This paper describes KANT, the first system to combine principled source language design, semi-automated knowledge acquisition, and knowledge compilation techniques to produce fast, high-quality translation to multiple languages.

1 Overview

Any expert system is only as good as the knowledge programmed into it; the same is true of a knowledge-based translation system. A KBMT system can only produce accurate, high-quality translations if it can unambiguously determine the meaning of the input text and choose an appropriate phrasing of that meaning in the target language. This implies a significant domain knowledge base in addition to the usual syntactic grammars, lexicons, etc. The question is, how much knowledge is enough?

It is probably the case that the implementation of a world knowledge base sufficient to support knowledge-based translation in any domain of discourse is some years from realization. Nevertheless, there are significant practical problems in translation that can be solved with current knowledge-based technology, because they do not require general translation in all domains of discourse. In particular, we have explored the use of machine translation for multi-language output in a controlled authoring environment for technical documentation.

Our goal has been to identify and develop the following:

- Principled designs for source language texts that encourage concise, expressive authoring while supporting efficient translation;
- Algorithms for knowledge-based interpretation of text that effectively disambiguate source language sentences;
- A powerful yet efficient rule formalism for target text generation;
- A combination of automatic and semi-automatic knowledge acquisition tools to streamline the creation of large-scale knowledge bases for translation in a particular application domain;
- Techniques for the compilation of knowledge bases that support a clear, declarative style of input for the linguist/knowledge engineer and produce efficient run-time knowledge sources for translation;
- A modular system architecture that allows extension to additional target languages without any change to existing knowledge.

In this project note, we describe the development of these ideas in the KANT system, a knowledge-based interlingua translation system for multi-lingual document production. We believe that KANT is the first system to bring these ideas together in a system that provides fast, accurate, high-quality knowledge-based translation.

A complete working prototype of the KANT architecture which translates to French, German, and Japanese has been demonstrated successfully, and KANT is currently being extended in a large-scale commercial application for document production in several languages.

1.1 Controlled Input Language

There are two broad classes of restrictions which KANT places on the source text. The first concerns the vocabulary used by the author. The general (non-domain specific) words used in the source text are limited to a basic vocabulary of about 14,000 distinct word senses. The domain-specific technical terms are limited to a pre-defined vocabulary. The second restriction concerns the level of syntactic complexity present in the source text. KANT limits the use of constructions that would create unnecessary ambiguity or other difficulties in parsing, while still providing the author with a subset of English which is large enough to support authoring of clear, understandable technical prose. For example, KANT allows the use of subject-gap relative clauses with an explicit relative pronoun (e.g., "Clean the ventilation slots *which are located on the rear of the chassis*"), but does not allow reduced relative clauses. An example of a controlled input language text is shown in Figure 2.

Previous attempts to define controlled input languages for translation have tried to reduce complexity by either limiting the vocabulary to a very small size or by limiting syntax to just a few constructions[1]. In contrast to systems which limit vocabulary to just a few thousand words, KANT allows a larger vocabulary to be represented in the lexicon. KANT also places principled grammatical limitations on the source text that are loose enough to allow a degree of stylistic variation which supports productive authoring, while controlling the complexity of the input in areas that are crucial for accurate translation.

1.2 Knowledge-Based Parsing and Interpretation

Although it is possible to reduce ambiguity by limiting the use of certain kinds of phrases, some phrases which introduce a high level of ambiguity (such as prepositional phrases) cannot be ruled out. To resolve the ambiguity introduced by multiple possible phrase attachments, KANT uses an explicit domain model to narrow the set of potential interpretations (cf. Figure 1). For every phrase (such as verb phrase or noun phrase) that accepts a potentially ambiguous phrase attachment (such as a prepositional phrase), KANT constrains the set of allowable attached phrases to just those that meet the narrow semantic restrictions of the particular domain. The system's domain model is rich enough to allow all interpretations possible within the domain, but narrow enough to rule out irrelevant interpretations. The complexity of the domain model is only as deep as required to resolve ambiguity, which is the appropriate criterion for limiting the size of a domain model in a practical KBMT system.

By constraining the set of possible syntactic structures and ruling out ambiguous interpretations, it is possible for KANT to assign a complete and accurate semantic representation to each input sentence. Although the creation of a comprehensive set of mapping rules requires intensive development, we have eliminated redundancy through structure-sharing and pre-compilation (Mitamura, 1989; Mitamura and Nyberg, 1990). Interpretation rules are organized into an inheritance hierarchy, so that general rules can be shared via inheritance; the hierarchy is then pre-compiled into cached structures for fast access at run-time.

1.3 A Powerful Rule Formalism for Generation

High-quality output in an Interlingua-based system presupposes a generation component that is powerful and flexible, allowing the system to create accurate target text realizations which do not necessarily reflect the syntactic organization of the source text or the structure of the Interlingua Text. The Mapper module of the system makes use of a set of mapping rules and a lexicon to create the appropriate Target F-Structure for each Interlingua representation (cf. Figure 1). Each mapping rule is intended to apply to a single Interlingua concept, which may contain other Interlingua concepts as slot fillers; the Mapper uses a recursive-descent f-structure composition algorithm, which is discussed in (Nyberg et al., 1991).

A mapping rule combines three types of information: a *pattern* slot, a context that must match the Interlingua concept to be mapped; a *syn* slot, a pointer to the lexical item to be used to realize the concept; and a *map* slot, which specifies how the embedded components of the Interlingua map to grammatical functions in the Target F-Structure. For example, the following rule maps the Interlingua concept `*E-REMOVE` to the French verb *déposer* in the appropriate context:

```
(glex *remove
  (pattern
     (theme (*or* *o-frame *o-chassis)))
  (syn
     (cat verb)
     (root "déposer"))
  (map (theme obj)))
```

The English sentence *Remove the chassis* would be translated to *Déposer le châssis* using this rule.

1.4 Automated and Semi-Automated Tools for Knowledge Acquisition

Since knowledge-based translation systems rely on the use of complex knowledge sources, knowledge acquisition becomes the single most important (and time-consuming) task during system development. The system must provide the developer with an efficient way to specify and incrementally refine both domain knowledge and linguistic knowledge. In addition, those parts of the development process that are most repetitive (such as the extraction of vocabulary lists from a text corpus) should be automated. The tools that are currently being used in the development of KANT applications include:

- *Structured Tools for Editing Domain Knowledge Sources.* We use the ONTOS knowledge acquisition tool, developed at the Center for Machine Translation, for the creation and update of our domain model (Kaufmann, 1991). ONTOS incorporates a graphic browser interface for rapid access with an integrated, structured editor to support development of large-scale domain hierarchies (Carlson and Nirenburg, 1990).
- *Automatic Corpus Analysis Tools.* To analyze quickly sample corpora for a domain under development, KANT makes use of automatic corpus analysis tools that segment the text and pre-process it to produce preliminary vocabulary lists. The tagged corpora are then available for selective on-line development and debugging of linguistic knowledge sources.

[1] For example, the Multinational Customized English used by XEROX Corporation (De Mauro and Russo, 1984) helped to decrease post-editing to the point where semi-automated translation became 5 times faster than manual translation.

- *Semi-Automated Acquisition Tools.* Following corpus analysis, KANT automatically extracts a syntactic lexicon and set of interpretation rules for the sample corpus. This is achieved by extracting the relevant vocabulary items from a master lexicon, and using a pre-defined mapping rule hierarchy and default mapping rule templates. These knowledge sources are then incrementally refined by the system developer once the bulk of the tedious work has been done automatically.

We are currently extending our tools so that they may be used to partially automate the process of knowledge acquisition for generation lexicons, grammars and mapping rules. We anticipate that this should not be difficult, since the formalisms used for generation knowledge are similar to those used in analysis.

1.5 Knowledge Pre-compilation for Run-time Efficiency

Knowledge-based translation systems require the use of several complex knowledge sources (e.g., grammars, mapping rules, domain models, etc.). It is important to support the declarative specification of knowledge sources to facilitate knowledge acquisition by human experts; on the other hand, it is absolutely necessary to encode that knowledge at run-time in the most efficient procedural form possible. Our system uses the Generalized LR Parser-Compiler (Tomita et al., 1988) to compile the LFG source grammar into a fast, efficient run-time parsing table. The GenKit grammar compiler (Tomita & Nyberg, 1988) is used to compile the LFG target grammar into a set of efficient CommonLisp functions for generation, which are further compiled into object code by the CommonLisp compiler. Our analysis and generation mapping rules are compiled into decision trees which optimize the amount of processing required to locate and evaluate the most appropriate mapping rule for a given syntactic structure or Interlingua concept. Although these compilation techniques have afforded us a high degree of run-time efficiency and acceptable translation speed, we are currently investigating the cross-compilation of our system into C to achieve further speed-up.

1.6 Modular System Architecture

To support efficient development of multi-lingual translation capability, KANT has a modular system architecture. The parser and generator are independent components (see Figure 1); as a result, any source language supported by the system can be translated to any target language supported by the system. This architecture allows knowledge sources for different languages to be combined easily in new applications to support various source and target combinations. It is also the case that a modular design decreases development time, since it allows parallel development of knowledge bases for the source and target language(s).

Each linguistic processing module in our system consists of a procedural and a declarative component, the procedural component capturing the general algorithm to be used, and the declarative component representing the specific knowledge required by that algorithm for a particular language. This makes it possible to add new knowledge for additional languages without having to re-write the general code for the system modules.

2 Characteristics of the KANT System

The KANT architecture has the following characteristics:

- **Semantic Accuracy and Completeness.**
  To be semantically accurate, a system must produce a complete, correct and unambiguous Interlingua representation for each input sentence; it must also produce a complete, correct and unambiguous output sentence for each Interlingua representation. In a narrow technical domain, KANT achieves near-perfect semantic accuracy. Once all relevant domain knowledge has been acquired by the system, the Interpreter is able to disambiguate any potentially ambiguous structural attachments to remove spurious interpretations of the input. The Interpreter also discards any Interlingua representations which are not complete interpretations of the Source F-Structure.

- **Grammatical Accuracy.**
  To achieve the objective of no post-editing, semantic accuracy by itself does not suffice. Accurate Interlingua representations cannot be produced unless the system has an adequate grasp of the source language syntax; nor can the system produce accurate target text from an accurate Interlingua unless it has adequate coverage of the target language syntax. In addition to purely semantic information, the Interlingua must also represent certain features of the input text, such as modality, aspect, discourse markers, etc. in order to generate grammatically accurate output texts. Our system uses explicit syntactic grammars, written in the LFG grammatical formalism, for the source language and target language(s). Our grammars include rules to handle both the basic sentential syntax of the language and discourse-level markers.

- **High Quality Output.**
  To go beyond semantic and grammatical accuracy and produce stylistically correct output, a translation system must have a good grasp of the textual structure of the target language as well as its sentential syntax. This requires an explicit representation of textual relations between clauses and sentences, and the ability to select and produce complex sentence structures when appropriate. The mapping rules used by KANT's Mapper can not only select the correct single phrase for an Interlingua concept, but also create more complex syntactic constructions when appropriate. Thus the ability of the system to generate stylistically correct output is limited only by the amount of effort dedicated to the construction of mapping rules for the target language.

3 Current Results

The present KANT prototype produces very accurate translations, without human disambiguation or post-editing, such as those illustrated in Figures 3-5. The system has been tested on a corpus of several hundred sentences of pre-authored text, with 100% accuracy and good quality. We intend to extend incrementally the coverage of KANT, while simultaneously maintaining the current level of accuracy and speed, in order to provide a smooth transition path from prototype to a larger-scale application system.

- The KANT prototype has been implemented in the domain of technical electronics manuals, and translates from English to Japanese, French and German.
- The current English lexicon contains about 14,000 general word senses and several hundred technical terms. The target language lexicons contain these technical terms and a smaller subset of the general terms, and are currently being extended. The eventual goal is to support a lexicon of 30-40,000 terms.
- The current Domain Model contains over 500 concept frames, which correspond to the meanings present in the sample corpora currently translated. We expect the size of the Domain Model to grow rapidly as more knowledge is acquired.
- KANT is implemented in CMU CommonLisp, and runs on IBM APC/RT workstations, which are rated at about 2.5 MIPS. Using this hardware, our system has achieved a translation speed of 1-3 seconds per sentence Faster translations are expected with newer hardware.

References

[1] Carlson, L. and S. Nirenburg (1990). *World Modelling for NLP*, Technical Report CMU-CMT-90-121, Center for Machine Translation, Carnegie Mellon University.

[2] DeMauro, P. and M. J. Russo (1984). "Computer Assisted Translation at XEROX Corporation," *Proceedings of the 25th Annual Conference of the American Translators Association*, New York, NY, September 19-23.

[3] Kaufmann, T. (1991). *The ONTOS User's Guide*. Technical Memo, Center for Machine Translation, Carnegie Mellon University.

[4] Mitamura, T. (1989). *The Hierarchical Organization of Predicate Frames for Interpretive Mapping in Natural Language Processing*, PhD thesis, University of Pittsburgh.

[5] Mitamura, T. and E. Nyberg (1990). "Multiple Inheritance and Interpretive Mapping in Machine Translation," unpublished manuscript.

[6] Nyberg, E., R. McCardell, D. Gates and S. Nirenburg (1991). "Target Text Generation," in Goodman and Nirenburg (eds), *A Case Study in Knowledge-Based Machine Translation*, San Mateo, CA: Morgan Kaufmann.

[7] Tomita, M., T. Mitamura, H. Musha, and M. Kee (1988). *The LR Parser-Compiler User's Guide, Version 8.1*, CMU-CMT-88-MEMO, Center for Machine Translation, Carnegie Mellon University.

[8] Tomita, M. and E. Nyberg (1988). *The GenKit and Transformation Kit User's Guide*, Technical Memo, Center for Machine Translation, Carnegie Mellon University, CMU-CMT-88-MEMO.

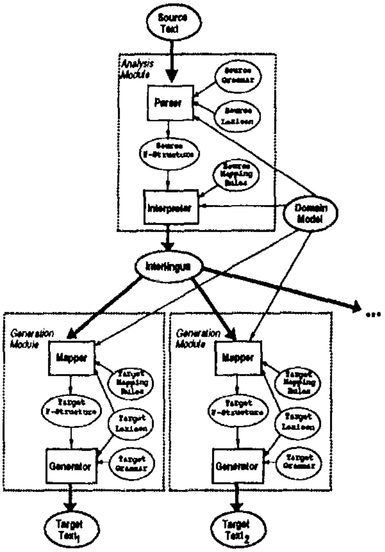


Figure 1: **KANT: Knowledge-Based Natural Language Translation**

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.

Figure 2: **Sample English Source Text Input to KANT**

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.

Figure 3: **French Target Text Produced by KANT**

Sicherheitsbestimmungen

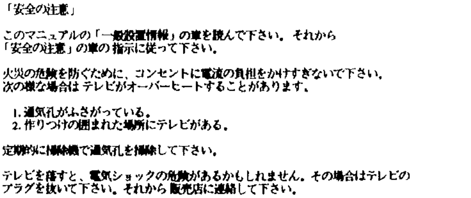
Lesen Sie den Abschnitt "Allgemeine Informationen zur Installation" in diesem Handbuch. Folgen Sie dann den Anweisungen in dem Abschnitt "Sicherheitsbestimmungen".
Vermeiden Sie Feuergefahren, indem Sie die Netzanschlüsse nicht überlasten.
Fernsehgeräte können in den folgenden Fällen überhitzen:

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.

Figure 4: **German Target Text Produced by KANT**



「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。

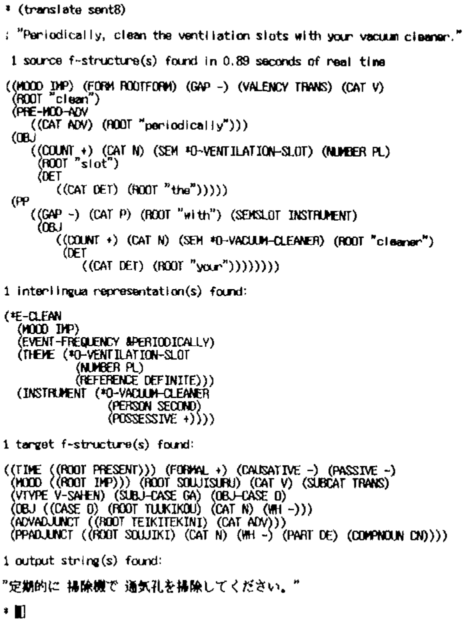
火災の危険を防ぐために、コンセントに電流の負担をかけすぎないで下さい。
次の様な場合は テレビがオーバーヒートすることがあります。

1. 通気孔がふさがっている。
2. 作りつけの囲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。

Figure 5: **Japanese Target Text Produced by KANT**



```
* (translate sent8)

; "Periodically, clean the ventilation slots with your vacuum cleaner."

 1 source f-structure(s) found in 0.89 seconds of real time

((MOOD IMP) (FORM ROOTFORM) (GAP -) (VALENCY TRANS) (CAT V)
 (ROOT "clean")
 (PRE-MOD-ADV
    ((CAT ADV) (ROOT "periodically")))
 (OBJ
    ((COUNT +) (CAT N) (SEM *O-VENTILATION-SLOT) (NUMBER PL)
     (ROOT "slot")
     (DET
        ((CAT DET) (ROOT "the")))))
 (PP
    ((GAP -) (CAT P) (ROOT "with") (SEMSLOT INSTRUMENT)
     (OBJ
        ((COUNT +) (CAT N) (SEM *O-VACUUM-CLEANER) (ROOT "cleaner")
         (DET
            ((CAT DET) (ROOT "your"))))))))

1 interlingua representation(s) found:

(*E-CLEAN
  (MOOD IMP)
  (EVENT-FREQUENCY &PERIODICALLY)
  (THEME (*O-VENTILATION-SLOT
           (NUMBER PL)
           (REFERENCE DEFINITE)))
  (INSTRUMENT (*O-VACUUM-CLEANER
                (PERSON SECOND)
                (POSSESSIVE +))))

1 target f-structure(s) found:

((TIME ((ROOT PRESENT))) (FORMAL +) (CAUSATIVE -) (PASSIVE -)
 (MOOD ((ROOT IMP))) (ROOT SOUJISURU) (CAT V) (SUBCAT TRANS)
 (VTYPE V-SAHEN) (SUBJ-CASE GA) (OBJ-CASE O)
 (OBJ ((CASE O) (ROOT TUUKIKOU) (CAT N) (WH -)))
 (ADVADJUNCT ((ROOT TEIKITEKINI) (CAT ADV)))
 (PPADJUNCT ((ROOT SOUJIKI) (CAT N) (WH -) (PART DE) (COMPNOUN CN))))

1 output string(s) found:

"定期的に 掃除機で 通気孔を掃除してください。"

* █
```

Figure 6: **Sample Translation to Japanese of One Selected Sentence, Showing Intermediate F-Structures and Interlingua Representation**