A REASONED INTERLINGUA FOR KNOWLEDGE-BASED MACHINE TRANSLATION

John R. R. Leavitt, Deryle W. Lonsdale, and Alexander M. Franz
Center for Machine Translation, Carnegie Mellon University,
Pittsburgh, Pa., USA, 15213.
jrrl@cs.cmu.edu, lonz@cs.cmu.edu, amf@cs.cmu.edu

Abstract

Research in machine translation (MT) has resulted in a number of Machine Translation systems that are based on the interlingua approach. This paper reports the results of designing an interlingua for a large-scale, practical MT system designed to translate technical information from English into a number of target languages. After an analysis of the main features of the translation problem faced by this system we describe the principles underlying our design decisions. We address issues such as the design and development methodology, the grain size of the representation, and our efforts to endow the interlingua representation with the ability to degrade gracefully. We conclude that large-scale MT of technical documentation can be achieved with an interlingua-based architecture if the pivotal point of the system, the interlingua, has been designed in a systems-oriented manner.

Keywords: Machine Translation, Applications, Knowledge Representation

1 Introduction

An interlingual architecture for machine translation (MT) has a number of advantages over other possible architectures, such as the "transfer" model. In an interlingua-based architecture, source text analysis and target text generation are divided into two separate components. An intermediate knowledge representation level, called the *interlingua*, mediates between the analysis and generation components. That is, the analysis component creates interlingua representations for the source text, and the generation component starts from the interlingua representation and creates target text from it. This allows the various knowledge sources (including knowledge sources for different languages) to be developed in parallel, and it creates an independence between the different components that greatly supports the development of more than one language pair.

On the other hand, an interlingua-based architecture creates a number of new difficulties, because the interlingua becomes the central pivot point for the entire translation process. This paper describes the design and implementation of the interlingua for the KANT project [Nyberg and Mitamura, 1992], a large-scale Machine Translation project for translating technical texts into multiple target languages.

2 Problem Definition

This section describes the problem of high-quality automatic translation of technical information.

2.1 Knowledge Sources

Machine translation requires the use of various types of knowledge (hence the term "Knowledge-Based Machine Translation" [Nirenburg et al., 1992]). For each language, this includes spelling, contraction, and formatting rules; morphological rules; lexical knowledge, including syntactic features, semantic concepts, and collocational and terminological information; knowledge about the grammatical structure; and semantic rules. In addition, a certain amount of knowledge about the domain is required.

2.2 Multiple Targets

The KANT project deals with machine translation as a tool for global information dissemination. For example, one KANT application under development translates technical service information for Caterpillar, Inc. products (heavy machinery) from English into the languages of the major export markets. Another example of a possible domain is user information for consumer electronics, such as television sets. Since there is only one source language, the analysis component can lean towards slight language dependence, but it is necessary to handle generation in multiple languages.

2.3 Technical Sublanguage Translation

KANT is a *sublanguage* translation system. That is, it is not designed to translate all of the English language, but rather a well-defined subset. An application sublanguage is constrained both by the domain from which the source texts are drawn (e.g. service information for heavy machinery), and by general restrictions that form a "Constrained Technical English." Since these restrictions define lexical, syntactic, semantic, and conceptual inventory that is in a fundamental sense closed (while of course remaining open for extensions within the framework), it is possible to achieve complete coverage for the source sublanguage during system development.

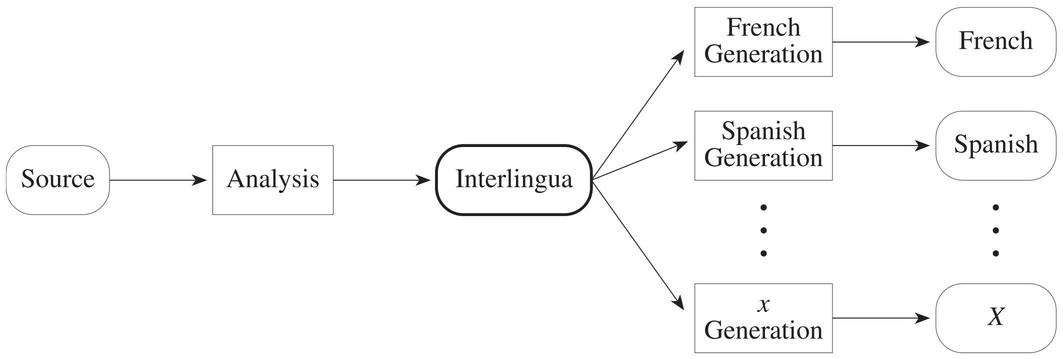

Figure 1: The Interlingual MT Architecture

Furthermore, Standard Generalized Markup Language (SGML) text markup codes are used in the input. Each application uses a tailored Document Type Definition (DTD) that includes tags for the logical and semantic structure of the domain. These tags are used directly during analysis, help to structure the source text, and are explicitly represented in the interlingua.

Compared to literary prose or poetry, technical information is conceptually, semantically, and pragmatically rather uncomplicated. This allows for another important feature: KANT translates on a sentence-per-sentence basis, and does not attempt to compute complicated pragmatic and discourse meanings. This means that the information that has to be represented in the interlingua is restricted to a feasible amount.

2.4 Modular Design

Multilingual MT is a complex problem. The different knowledge sources for the various languages need to be developed by separate language experts, and domain knowledge has to be encoded into the domain model by an expert in the domain. This calls for a modular architecture that separates knowledge sources from processing engines, shields different languages from each other to avoid language-pair-specific development, and that provides module interfaces that are habitable for language experts who are not necessarily skilled programmers.

2.5 The Interlingua Approach

The solution to these challenges is to divide the problem into source language analysis, and target language generation (see Figure 1). The interface between these two components is an intermediate language called the *interlingua*. It is a language-independent, unambiguous representation of the meaning of the input text that has to fulfill a simple functional condition: the interlingua representation must be sufficient for accurate translation in a technical domain.

3 Design Principles

This section describes the principles underlying the design of the KANT interlingua. Section 4 shows how these principles were put into practice.

3.1 Reasoned Approach

Some MT projects adopt a certain theory or methodology at the onset and adhere to it throughout development. In our approach, the first step is to analyze the problem, and then to "reason out" a design that results in a practical, working system.

3.2 Staged Approach

Since the interlingua plays a central role in the system, we chose to develop it by a method of iterative refinement. This allows development of the separate components to proceed smoothly in parallel not only with each other, but with the development of the interlingua itself.

3.3 Balanced Approach

[Tsujii, 1988] discusses three general ways to approach the task of choosing and implementing an interlingua:

- top-down: by considering the domain and enumerating *a priori* the concepts, processes, and relationships required for its treatment;
- bottom-up: by considering the (disambiguated) lexical content expressed by text discussing the domain, and then defining sets, hierarchies, and other relevant relationships; and
- decompositional: by (re-)expressing all relevant aspects of the domain with respect to a highly restricted set of semantic primitives.

We believe that the best design method is a combination of the top-down and bottom-up approaches. Previous work, and linguistic theory, guides an initial top-down structuring of the domain of the interlingua. Then, corpus-based incremental bottom-up work extends the interlingua towards complete coverage of the domain.

3.4 Comprehensive Approach

We believe that the interlingua must represent information from all necessary levels of linguistic analysis: Lexical, syntactic, and pragmatic. The interlingua is designed to represent all such necessary features, but no more.

4 The KANT Interlingua

The KANT interlingua is a recursive representation scheme using nested frames to correspond to information contents of elements of source sentences. Each interlingua frame contains a conceptual head and a series of feature-value pairs and semantic slots which in turn contain additional interlingua frames. Concepts can correspond to source language expressions (e.g. the action `*a-bond`), semantic units from the domain (e.g. `*c-decimal-number`), or structural elements such as tagged SGML constituents. A sample interlingua structure from the domain of heavy equipment manufacturing is shown in Figure 2.

Space prohibits a complete explanation of all the parts of the KANT interlingua here, but it is described extensively in [Leavitt et al., 1993]. This document includes not only detailed discussions of the semantic roles, concepts, and features used in interlingua, but also extensive examples of their usage, which also serve as a test suite for the implementation.

"The gasket must be bonded to the valve cover between the bolt holes (14 places) with <product> permabond 102 <\product>." $\Longrightarrow$

```
(*a-bond
   (punctuation period)
   (mood declarative)
   (obligation medium)
   (tense present)
   (topic-role theme)
   (theme (*o-gasket
            (number singular)
            (reference definite)))
   (attach_to (*o-valve-cover
                (number singular)
                (reference definite)))
   (located_between
      (*o-bolt-hole
         (number plural)
         (reference definite)
         (parenthetical
            (*o-place
               (number plural)
               (generic +)
               (quantity
                  (*c-decimal-number
                     (integer "14")
                     (number-type cardinal)
                     (number-form numeric)))))))
   (means_with (*s-product
                  (value "Permabond 102"))))
```

Figure 2: Example KANT Interlingua Structure

4.1 Conceptual Grain Size

Past MT efforts have addressed the problem of specificity in interlingual representations in several different ways. Many involve the creation of a highly structured language-neutral representation which will not be subject to the nuances, ambiguities, and other vagaries of natural linguistic usage. In addressing the granularity issue for the representation, we chose to combine the best features from different possible design paradigms.

With respect to the choice of primitives, we proceeded in both bottom-up and top-down fashions. On the one hand, we established a circumscription of the domain addressed, discourse styles, and typical document structure. At the same time, though, we carried out extensive bottom-up identification of the domain through extraction of knowledge by automated corpus analysis techniques.

From a linguistic perspective, we also proceeded in both directions. Whereas others have sought to incorporate into an interlingua framework the expressive machinery of formalisms from logic, programming languages, or descriptive linguistic theories, direct interpretation cannot always guarantee an appropriate granularity for a unified and comprehensive treatment of linguistic phenomena. We have chosen to avoid the small grain size of interlingual approaches like UNITRAN, which is based on a theory of lexical semantic description. While useful in other contexts, and in fact are not incompatible with a KBMT approach, it seemed overly specific for our needs.

To illustrate this overspecificity, we include examples from two other interlinguas. Figure 3 shows some excerpts from the TAMERLAN representation for an advertising text for a doughnut store[Nirenburg and Defrise, 1994]. The representation covers more than individual sentences, and it is structured as a sequence of frames that are related to each other via pointers. As can be seen from the example, TAMERLAN represents a far greater level of detail, and includes pragmatic and discourse information.

Dorr [1993] uses a different, but still overly rich, type of interlingua. An example of this structure is shown in Figure 4.

We have found that rather that attempting to attain the most diminutive grain size, it is quite useful to perform lexical chunking, combining such highly lexicalized items as fixed phrases, technical nomenclature, and company-specific terminology.

Our grain-size was also set based on an examination of the communication content of the documents we address. Realizing that the complexity of natural language phenomena tends to favor the creation of complicated interlingua structures, we sought to avoid too complex a set of relations, features, and concepts for our constrained domain. For example, some systems seek to establish a comprehensive modeling from the perspective of an intelligent agent interacting with a complex environment, and following goal-directed behavior involving interpretation of discourse and situational contexts (cf. [Defrise, 1993]). Certain types of text may well require such considerations; for ours, they tend to introduce avoidable overhead and complexity.

Clearly our goal was to design a minimalist representation, considering both the breadth and depth of the domain addressed, and avoiding the opposing pitfalls of over-complexity and under-specificity.

"Drop by your old favorite Dunkin' Donuts shop . . . " $\Longrightarrow$

```
(make_frame text_1
  (clauses
    (value clause_1 clause_2 clause_3 ...))
  (relations
    (value relation_1 relation_2 ...))
  (attitudes
    (value attitude_1 attitude_2 ...))
  (producer-intentions
    (value producer-intention_1)))

(make_frame clause_1
  (head (value %visit_1))
  (aspect
    (duration prolonged)
    (phase begin)
    (iteration 1))
  (time (value time_2)))

(make_frame %visit_1
  (is-token-of (value *visit))
  (agent (value *consumer*))
  (destination (value %shop_1)))
                              ⋮
```

Figure 3: Example Tamerlan Interlingua Structure

4.2 Graceful degradation

In that the KANT interlingua is a simple recursive data structure, it contains very little interrelation between elements. Only the interrelations that are necessary for generating output are represented. Furthermore, it is possible for the generation component to fail to realize various portions of the interlingua and still produce acceptable (if incomplete) output. For example, consider the interlingua structure in Figure 2. During generation, if the contents of the `means_with` or `parenthetical` slot were not realized, the resulting sentence would still convey the main idea of the sentence.

Similarly, if an incorrect value is assigned to a feature, an incorrect, but still comprehesible sentence can usually be produced. For example, during the early development of the system, the values of the `number` feature for objects were changed from `SG` and `PL` to `Singular` and `Plural`. For a while after that decision was made, an occasional `(number sg)` or `(number pl)` would appear. In these cases, the generator would usually produce a sentence in which the only error was the number on the noun phrase produced from the object and perhaps the subject-verb agreement if the affected object was mapped to the subject. Otherwise, the sentences were just fine and were certainly comprehensible.

Of course, there are places where a missed feature or semantic role would cause less graceful degradation. For example, failing to realize the contents of the `attach_to` or `theme` slot in the example interlingua structure would produce unacceptable output. There are, in essence, two classes of information in the interlingua — that which is necessary to create acceptable output and that which is not. If the latter is missing, the KANT interlingua may still maintain integrity, while an error or omission within the first class will cause a legitimate failure. Our interlingua differs from others in that we have tried to shift as much information as possible from the former class into the latter, by minimizing interactions between elements.

"I stabbed John." $\Longrightarrow$

$$
\begin{array}{l}
[_{\text{EVENT}}\ \text{CAUSE}\ ([_{\text{THING}}\ \text{I}], \\
\quad [_{\text{EVENT}}\ \text{GO}_{\text{Poss}}\ ([_{\text{THING}}\text{KNIFE-WOUND}], \\
\quad\quad [_{\text{PATH}}\ \text{TOWARD}_{\text{Poss}} \\
\quad\quad\quad ([_{\text{POSITION}}\ \text{AT}_{\text{Poss}} \\
\quad\quad\quad\quad ([_{\text{THING}}\ \text{KNIFE-WOUND}],\ [_{\text{THING}}\ \text{JOHN}])])])])]
\end{array}
$$

Figure 4: Example Dorr [1993] Interlingua Structure

In addition, since the structures are simple, generalized rules can be written for the generation components to handle most constructions, which further minimizes the chance of poor output. If the information for a given sentence were represented as a collection of objects that are connected only by a number of highly interreactive links, this would be more difficult.

Similarly, on the analysis side, if a semantic role cannot be determined for a given piece of information, a generic role may be used instead (i.e. `generic_with` instead of `goal_with`). Generalized rules for mapping these slots can be included on the generation side, which results in a further decrease in loss of information and quality.

4.3 Specification

Throughout the design, a specification document was maintained, to which both analysis and generation modules of the system could refer for the latest interlingua design.

After initial consideration of such issues as lexical precision, meaning preservation, syntactic markedness, metatextual reference, and semantic hierarchies, we established a set of relevant data types and representations. These data structure had to be habitable and mnemonic, since developers can find working with an interlingua difficult because of its high degree of abstraction away from the lexical form of language. Development of large-scale interlingua systems such as ours must not be further complicated by an opaque or obscure data format.

These decisions formed the basis of the specification document and test suite, both of which were iteratively refined.

4.4 Iterative Refinement

One difficulty with interlingual system is what has been the "horizon effect" seen in other endeavors of natural language processing. As work proceeds at one level of definition, specification, and implementation, a certain degree of rigor and expressiveness of the interlingua is attained. Still, however carefully this has been achieved, the expressive power of natural language, even in restricted subdomains, tends to always favor greater complexity. In effect a "new horizon" becomes visible, inviting a further increment to the interlingua development process.

This effect is minimized in two ways by our approach. First, the principle of parsimony combined with the nature of our target domains effectively limits the amount of information that the KANT interlingua needs to represent. There are certain phenomena (pragmatics factors, speaker intentions, discourse levels, etc.), which either are not significant in technical documentation or can be eliminated via rewriting. The KANT interlingua will not have to represent these phenomena and so certain extensions to the horizon become impossible. Second, due to the central role of the interlingua in the system, a rapid prototyping and incremental refinement strategy, similar to the spiral software development model [Boehm, 1985]), is necessary. By planning incremental refinement into the design process, the remaining horizon effect is transformed into a forcing function for each refinement iteration. In essence, the horizon effect becomes part of the design process, rather than a force opposing it.

There are also additional advantages of the incremental approach. As mentioned earlier, in following an incremental refinement strategy in designing the interlingua, both the analysis and generation components of the system could be developed in parallel not only with each other, but also with the development of the interlingua itself.

The interlingua is the totality of information passed by the analysis module to the generator. Since the interlingua plays this central role in the system, its design can easily become a bottleneck for the entire development process, as neither analysis nor generation can proceed without first having the interlingua specified. In addition, the relationships between the concepts represented in the interlingua also evolve incrementally as the related frame-based hierarchical domain model is refined.

By allowing for incremental development, the interlingua's central role in the system, rather than constituting a bottleneck, became for us a focal point of development effort. That is, both sides were able to be proceed using the latest information about the interlingua and provide feedback to the the interlingua design process.

For example, the semantic roles that correspond to English prepositional phrases were not specified until well into the interlingua design process. However, because the heart of the interlingua had already been specified, both the analysis and generation teams were able to work with this underspecified form (albeit without any source prepositional phrases being represented in the interlingua). Similarly, the domain model, which supplies the concepts used to head interlingua frames, was also developed in parallel. It was possible, when necessary to refine the grain size of the conceptual objects (e.g. to account for new subtleties of meaning) without affecting any other parts of the interlingua, and therefore without impacting more than necessary on the analysis and generation efforts.

A final advantage of adopting an incremental refinement strategy from the onset is that information obtained during testing may be used as design feedback. It is not uncommon for many representation problems to go unnoticed until actual text is fed through the MT system. In many cases, this form of feedback can be difficult to integrate back into the interlingual design, because the design and the design methodologies do not support iterative development. In our system, we were able to use the results of testing to create a better interlingua specification, rather than having to retrofit new structure into a existing inflexible base.

While in theory this approach does not eliminate the development bottleneck since the software development could overtake the interlingua development, in practice this is unlikely.

5 Conclusion

In this paper we have identified and described the approach that we have followed in the design and implementation of the interlingua for the KANT knowledge-based MT system. We have shown how, in order to achieve a large-scale practical system, fundamental software research and development principles must be followed. Our experience indicates that such efforts are only possible when the central knowledge representation is sufficiently expressive yet constrained, thorough yet practical, and well-specified yet extensible.

The reasoned approach we describe has been validated by the KANT application discussed, the first instantiation of which is about to be deployed at Caterpillar, Inc. for translation from English to French. The approach has been tested for Japanese, German, and Italian generation and development for large-scale Spanish and German generation components is already underway with additional languages to follow.

6 Acknowledgments

We would like to acknowledge the other members of the KANT team, in particular Jaime Carbonell, Eric Nyberg, Teruko Mitamura, Kathy Baker, Marion Kee, and William Walker. We also appreciate the collaboration of our associates at Carnegie Group Inc. and Caterpillar Inc.

References

[Boehm, 1985] Boehm, B. (1985). A spiral model of software development and enhancement. In *Proceedings of the International Workshop on Software Process and Software Environments*.

[Defrise, 1993] Defrise, C. (1993). Discours et traduction automatique: une approche interlangue basée sur les connaissances. In Bouillon, P. and Clas, A., editors, *La traductique*. Les Presses de l'Université de Montréal.

[Leavitt et al., 1993] Leavitt, J., Franz, A., and Lonsdale, D. (1993). The KANT interlingua specification. Technical Report CMU-CMT-93-143, Center for Machine Translation, Carnegie Mellon University.

[Nirenburg et al., 1992] Nirenburg, S., Carbonell, J., Tomita, M., and Goodman, K. (1992). *Machine Translation: A Knowledge-based Approach*. Morgan Kaufman, San Mateo, CA.

[Nirenburg and Defrise, 1994] Nirenburg, S. and Defrise, C. (1994). Application-oriented computation semantics. In Johnson, R. and Rosner, M., editors, *Computational Linguistics and Formal Semantics*. Cambridge University Press.

[Nyberg and Mitamura, 1992] Nyberg, E. and Mitamura, T. (1992). The KANT system: Fast, accurate, high-quality translation in practical domains. In *Coling-92*.

[Tsujii, 1988] Tsujii, J. (1988). What is a cross-linguistically valid interpretation of discourse? In Maxwell, D., Schubert, K., and Witkam, A. P. M., editors, *New Directions in Machine Translation*. Foris Publishers.