LARGE-SCALE MACHINE TRANSLATION: AN INTERLINGUA APPROACH

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

Abstract

In this paper we discuss the design and development of an interlingua for a large-scale MT project. We also discuss how the resulting KANT interlingua constrains complexity, supports staged development, evolves in a balanced fashion, and seeks maximal coverage. We address issues such as granularity of the data representation, its specification, the types of information it encodes, and how it supports a modular system architecture. Our experience shows that through a careful, reasoned design and implementation effort, it is possible to achieve multilingual target-language generation for extensive technical domains.

Keywords: Natural Language Processing, Knowledge Representation, Practical Applications

1 Introduction

Varying degrees of success have been achieved by researchers and developers of machine translation (MT) systems. Automatic machine translation of technical text is possible, if the problem is approached the right way. Systems using a knowledge-based approach are usually characterized by a highly modular architecture. This allows various knowledge sources to be developed in parallel, but places great demands on the interface between different languages. This paper describes the design of the *interlingua*, the intermediate knowledge representation that mediates between different languages. We focus on the requirements for translation of technical text, the main challenges, and the solutions that were adopted by the KANT project [Nyberg and Mitamura, 1992]. An idea of the scale and success of this large-scale undertaking is also given through informal discussion of the knowledge sources and domain size addressed.

2 Requirements

In an increasingly global trade community, translation has become a highly strategic and time-critical element in the development and marketing of many products. The efficient publication of high-quality documentation in languages other than English is often an essential activity in current manufacturing and development sectors of industry. Machine translation can assist in the efficient, consistent, and timely production of such documentation. This section describes the requirements for high-quality automatic translation of technical information. The KANT system we discuss translates technical service information for a wide range of Caterpillar, Inc. products (heavy machinery) from English into languages of the major export markets. Another example of a possible domain would be user information for consumer electronics.

2.1 The Knowledge Base

Successful translation by machine 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, collocational and terminological information; knowledge about grammatical structure; and semantic rules. In addition, a certain amount of real-world knowledge about the domain is required. The coverage attained by the current instantiation of the KANT system spans a domain consisting of over 60,000 concepts. The almost 2000 action concepts, derived from different verbal constructions, admit one or more of 35 different argument subcategorization frames (such as agent+theme, experiencer+stimulus, or agent+patient+attribute).

2.2 Multilingual Generation

The KANT project establishes machine translation as a tool for global information dissemination. As such, its central data structure, the interlingua, must be robust enough to support translation into several different target languages. It must reflect as closely as possible the content expressed in the original source text. It must not encourage any kind of ambiguity. The wide range of possible actions, properties, and objects, as well as the relationships between them, must be concisely captured.

In fact, over 500 source-language phrase-structure rules have been developed and are used to generate syntactic parses from the input.These parses are in turn interpreted and reformulated into interlingua structures.

Target language text generation must be able to use whatever semantic, syntactic, and lexical information is needed to reexpress appropriately the interlingua's conceptual content. Again, hundreds of rules are required to assure that the target output is morphologically, syntactically, and semantically well-formed.

Though the analysis component must support generation in multiple languages, it currently handles only one source language, and therefore can tolerate a slight degree of source language dependence.

2.3 Translating Technical Information

KANT is a *sublanguage* translation system. It is not designed to translate all possible text written in English, 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 a lexical, syntactic, semantic, and conceptual inventory that is in a fundamental sense closed (while remaining open for extensions within the framework), it is possible to achieve complete coverage for the source sublanguage during system development.

Furthermore, Standard Generalized Markup Language (SGML) text markup codes are used in the input. A KANT application uses a tailored Document Type Definition (DTD) that includes tags for the logical and semantic structure of the domain. These tags impose structure on the input text while demarcating certain of its salient features. Tags of linguistic significance are parsed and interpreted by the analysis module and then represented directly in the interlingua representation. The dozens of carefully designed tags allow consistent delimitation and handling of the customer's considerable documentation materials, product line nomenclature, and wide range of sophisticated measurement values and ranges. Only the information necessary for outputting such information in the target language is retained; the source language-specific conventions are not reflected in the interlingua.

The source language text is controlled with an interactive authoring environment which assists document authors and editors. It assures conformity to well-defined document production standards, covering all aspects including text markup, spelling and grammar usage, and even measurement formats.

Compared to literary prose or poetry, technical information is conceptually, semantically, and pragmatically quite uncomplicated. This allows 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 only a manageable amount of information has to be represented in the interlingua.

2.4 Managing Complexity

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 in 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 (avoiding language-pair-specific development), and provides module interfaces that are habitable for domain and language experts (who are not necessarily skilled programmers). System development must therefore involve minimizing the complexity of the task wherever possible.

It should be stressed that, while it is possible to constrain textual markup, lexical content, and syntactic structures to reduce complexity of the task, such a course must be taken carefully. Usability of the final system must not be compromised. The KANT system is capable of handling text which, though to a degree constrained, is nevertheless quite natural in style. In fact, it is often indistinguishable from standard, unconstrained materials. This protects the customer from undue hardship due to complexity issues.

2.5 Interlingual Machine Translation

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 Considerations

This section describes considerations underlying the design of the KANT interlingua. Section 4 shows how they were addressed.

3.1 Practicality

Some MT projects adopt a certain theory or methodology at the onset and adhere to it throughout development. While theoretically satisfying, such an enterprise is not always possible or appropriate in the development of a large-scale practical system. In our approach, the first step was to analyze the problem, and then to carry out an interlingua design that would lead to a practical, working system.

3.2 Incrementability

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 in parallel not only with each other, but also with the development of the interlingua itself. From the prototype stage to full-scale production deployment, various benchmarks were planned, reached, and evaluated.

3.3 Completeness

Our interlingua has emerged from a balanced approach following both empiricist and rationalist methods. A general discussion of interlingua construction can be found in [Tsujii, 1988], which mentions two of the methods we use.

On the one hand, we proceeded in *top-down* fashion by considering the domain and enumerating *a priori* the concepts, processes, and relationships required for its treatment. This involved considerable consultation with technical experts well versed in the domain addressed. In the context of our project, we were able to combine with the professional experience of the customer's experts a state-of-the-art theoretical soundness from recent academic MT research.

At the same time, we also carried out a *bottom-up* development strategy. Recognizing the importance of the corporate multilingual history captured by the significant corpus of already translated documentation, we were able to leverage this resource in order to determine the lexical inventory of the domain (see [Mitamura et al., 1993]).

A wide range of automated natural language processing tools including text deformatters, KWIC editors, multilingual

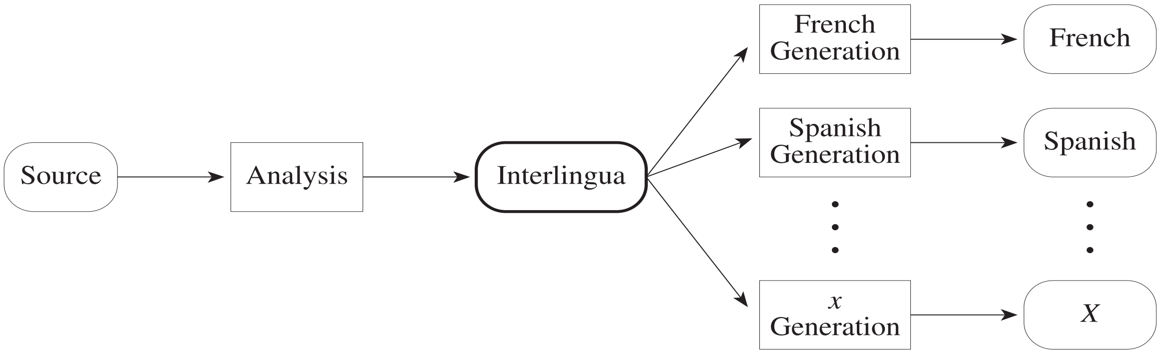

Figure 1: The Interlingual MT Architecture

corpus aligners, and translation extractors allowed domain experts to effectively review and annotate data that was required for domain and interlingua definition.

We found that the best design method was a combination of the top-down and bottom-up approaches. Domain experts and linguistic theory guided an initial top-down structuring of the domain addressed by the interlingua. In addition, corpus-based incremental bottom-up work extended the interlingua towards complete coverage of the domain.

3.4 Comprehensiveness

We believe that the interlingua must represent information from all necessary levels of linguistic analysis: lexical, syntactic, semantic, and pragmatic. The interlingua is designed to represent all such necessary features, but no more. To be useful, its constraints must still allow it to be selectively comprehensive.

4 Solutions

The KANT interlingua is a recursive list-based structural representation of the information content of individual source sentences. An interlingua frame consists of a head concept, feature-value pairs, and semantic slots which in turn contain nested interlingua frames. Concepts can correspond to source language expressions (e.g. the action `*a-supply`) or semantic units from the domain (e.g. `*c-decimal-number`). The overall format is modeled after the well-established tradition using frame-based structures reflecting deep semantic relationships between major constituents (see, for example, [Bruce, 1975]).

A sample interlingua structure from the domain of television repair manuals 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 the interlingua, but also extensive examples of their usage, which serve as a test suite for the implementation.

4.1 Representation Granularity

Past MT efforts have addressed the problem of specificity in interlingual representations in several different ways. Many

"The primary power supply component will supply the necessary 240 Volts DC to the input lead." $\Longrightarrow$

```
(*a-supply
   (tense future)
   (mood declarative)
   (punctuation period)
   (source (*o-power-supply-component
             (reference definite)
             (number singular)
             (attribute (*p-primary))))
   (theme (*u-volt-dc
            (reference definite)
            (number plural)
            (attribute (*p-necessary))
            (quantity
               (*c-decimal-number
               (integer "240")
               (number-type cardinal)
               (number-form numeric)))))
   (goal_to (*o-input-lead
               (reference definite)
               (number singular))))
```

Figure 2: Example Sentence and Interlingua Structure

projects involve the creation of a highly structured language-neutral representation which will not be subject to the nuances, ambiguities, and similar difficulties of natural linguistic usage. Alternatively, some have resorted to the use of human languages, planned or otherwise, such as Esperanto [Witkam, 1988] and even Aymara [Guzmán de Rojas, 1988], as interlinguas. Others (e.g. [Kittredge, 1993]) generate multiple target-language texts (like marine weather forecasts or employment statistics reports) directly from raw data interlinguas collected by expert systems over a subdomain, effectively sidestepping any natural language complexities on input. In addressing the granularity issue for the representation, we chose to combine the best features from different possible design paradigms.

As discussed above, our choice of primitives 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 not incompatible with a KBMT approach (see [Dorr, 1993]), it seemed overly specific for our needs. For example, we find it useful to perform lexical chunking, combining such highly lexicalized items as fixed phrases, technical nomenclature, and company-specific phraseology. These, in fact, constitute a large portion of a company's closely-guarded (sometimes even proprietary) terminological identity.

Our grain-size was also determined after an examination of the information 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 model 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.

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.

4.2 Graceful Degradation

The KANT interlingua is a simple recursive data structure and, as such, does not contain much 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, if the the contents of the `goal_to` slot of the interlingua structure in Figure 2 were not realized, the resulting sentence would still convey the main idea of the sentence.

Naturally, there are places where a missed feature or semantic role would cause less graceful degradation (failing to realize the contents of the `quantity` slot in the interlingua in Figure 2, for example). 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.

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_to` instead of `goal_to`). 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 period, a specification document was maintained, to which both analysis and generation efforts referred for the latest interlingua information.

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 Incremental Development

One difficulty with interlingua systems is what has been called the "horizon effect" seen in other endeavors of natural language processing and AI. As work proceeds at one level of definition, specification, and implementation, a certain degree of rigor and expressiveness of the interlingua is attained. Still, the expressive power of natural language, even in restricted subdomains, tends to favor greater complexity. In effect a "new horizon" becomes visible, inviting a further increment to the interlingua development process. Obviously unanticipated horizons can prove costly if not properly quantified and dealt with.

This effect is minimized in two ways by our approach. First, the principle of parsimony combined with the nature of our controlled source text and domains effectively limits the amount of information that the KANT interlingua needs to represent. There are certain phenomena (pragmatic factors, speaker intentions, discourse levels, etc.), which either are not significant in technical documentation or can be eliminated via rewriting. The KANT interlingua does not have to represent these phenomena and so certain extensions to the horizon become unnecessary. Second, due to the central role of the interlingua in the system, a rapid prototyping and incremental refinement strategy is necessary. By planning incremental refinement as part of 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 hindering 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 proceed using the latest information about the interlingua and provide feedback to the the interlingua design process. While in theory this does not eliminate the development bottleneck since the software development could overtake the interlingua development, in practice this is unlikely.

5 Challenges

While a KANT interlingua approach has proven successful for the large-scale system described, there are still notions which should be refined and reexamined. In this section we briefly mention some of these future challenges.

Given the vast range of possible input text types and the size of the domain addressed, it has proven difficult to quantify, identify, and select a body of text which can serve as a thorough and methodical corpus of test material. All aspects of analysis and generation must be identified, addressed, and exercised. While sizable test corpora have been assembled by developers for testing, the collection of an exhaustive ensemble of Controlled Technical English documents is not a trivial task.

We have experienced several iterations of the "horizon effect" described above while specifying concept grain sizes. For example, a single concept like `*o-bank` may have at first been associated with the denotations of the English word "bank". As development progressed, though, it would become apparent that the domain supported two different senses: "a mound or pile of dirt", and "a group of components arranged in a row", though not "a financial institution". In such cases we chose to refine the original concept, creating two more precise ones instead.

Another granularity issue is related to the wide use of prepositional phrases in English. Over 140 prepositions are allowed in the controlled language, and their meanings represent hundreds of possible distinct relationships between actions and/or objects. Whereas a handful of primitive prepositional role attachments was clearly inadequate for our purposes, our current set of several hundred could probably be reduced. For example, the primitive attachment `location` only vaguely represents the full meaning of several prepositions. We have instead refined this attachment to include such roles as `located-alongside` and `located-across`. The point of diminishing returns when such detailed attachments are used is not yet entirely clear.

Regarding the interlingua representation itself, our design has sought to express underlying associations between sentence constituents wherever possible and necessary. Though this allows us to handle such phenomena as scope for quantifiers, negation, and adverbials, our coverage is not entirely complete. We are still examining which combinations of feature-value pairs and distinct slot assignments are appropriate for satisfactory representation of these issues in the interlingua. We also face these questions in dealing with phenomena such as modality, aspect, and cross-category structures such as nominalizations.

Many of these challenges are well known to those who have undertaken development of sizable interlingua systems, and solutions are yet to be found. In our approach we have been able to benefit from the modularity and extendability that only an interlingua-based architecture can facilitate.

6 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 build a large-scale industrial 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 practical approach we describe has been validated by the KANT application which is about to be deployed at Caterpillar, Inc. for translation from English to French. Development of the Spanish and German generation components is already underway, and additional languages will follow.

7 Acknowledgments

We would like to acknowledge the other members of the CMT 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., Caterpillar Inc., and Traductions Taurus.

8 References

[Bruce, 1975] Bruce, B. (1975). Case systems for natural language. *Artificial Intelligence*, 6.

[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.

[Dorr, 1993] Dorr, B. J. (1992/1993). The use of lexical semantics in interlingual machine translation. *Machine Translation*, 4/3.

[Guzmán de Rojas, 1988] Guzmán de Rojas, I. (1988). ATAMIRI — interlingual MT using the Aymara language. In Maxwell, D., Schubert, K., and Witkam, A. P. M., editors, *New Directions in Machine Translation*. Foris Publishers.

[Kittredge, 1993] Kittredge, R. (1993). MT technology and text generation. In Nirenburg, S., editor, *Progress in Machine Translation*, pages 291–292. IOS Press.

[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.

[Mitamura et al., 1993] Mitamura, T., Nyberg, E., and Carbonell, J. (1993). Automated corpus analysis and the acquisition of large, multi-lingual knowledge bases for MT. In *5th International Conference on Theoretical and Methodological Issues in Machine Translation*, Kyoto, Japan.

[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.

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

[Witkam, 1988] Witkam, T. (1988). DLT - an industrial R&D project for multilingual MT. In *Proceedings of COLING-88*.