Can Practical Interlinguas Be Used for Difficult Analysis Problems?

Krzysztof Czuba
Teruko Mitamura
Eric Nyberg

Language Technologies Institute
Carnegie Mellon University

{kczuba,teruko,ehn}@cs.cmu.edu

1. Introduction

In this paper, we describe the results of an exercise in applying the source language analyzer from the KANT system (Mitamura, et al. 1991; Nyberg et al., 1998) to the English version of the workshop newspaper article. We briefly present the architecture of the KANT system and the Interlingua format used in KANT. We describe how we adapted the analyzer in order to produce an Interlingua for each sentence and discuss the issues we encountered and the resulting representations[1].

Since the KANT system has been designed for well-defined, technical domains and has been used mostly with controlled language input to obtain high-quality machine translation of technical documents, at the outset of this experiment we were not sure if the KANT Interlingua would be sufficiently sophisticated to provide usable representations of the sentences in the sample text for the purpose of MT. As the results of the experiment show, we were able to automatically generate an Interlingua for all sentences[2] in the sample text. In our opinion, the resulting representations would provide a good basis for the KANT generation module for languages like German, Spanish or French, although the representation difficulties we discuss below might make the Interlinguas less useful for other types of applications.

[1]See the last page of this paper for pointers to the full set of interlingua representations, which are available on the World-Wide Web.

[2]We did not analyze the titles of the parts of the article. In general, titles might require special treatment in both the analysis grammar and the Interlingua representation which currently are not part of our design in an uncontrolled environment.

2. Analyzer Architecture

The architecture of the KANT analyzer is illustrated in Figure 1. Using a lexicon and a unification grammar[3] a set of LFG-style f-structures (possible syntactic analyses) is produced. The Interpreter module then maps each f-structure to an Interlingua. The

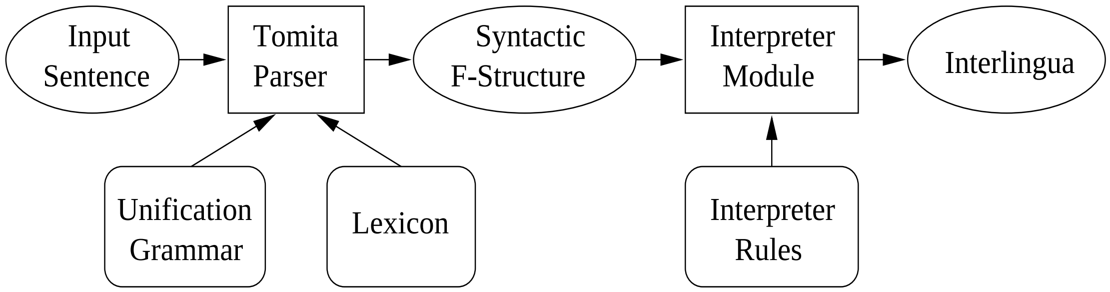

Figure 1: KANT Analyzer Architecture

Interpreter rules are responsible for assigning semantic concepts to each constituent in the sentence, mapping grammatical features to semantic features, and assigning semantic roles to various structural components of each sentence (verb arguments, PPs, etc.). The mapping of the SL f-structure to the Interlingua is done on the basis of lexical information associated with the syntactic lexical entries of the source language words, the syntactic information in the f-structure and the concepts appearing in the Interlingua[4].

In the KANT MT system, the Interlingua is then mapped to the target language f-structure. During mapping to the TL f-structure, the information contained in the Interlingua can be rearranged so that a representation closer to the target language syntax can be obtained. In particular, some reasoning concerning the target language argument structure, temporal relationships and structure of constituent modification is done at this point. This architecture design decision leads to a simpler Interlingua design at the cost of an extra procedural mapping step: although a pure Interlingua design would completely abstract away from the details of the source language, in a practical MT application, it is often easier to allow such details to influence the Interlingua structure. It is also desirable in some cases as it allows for a more precise translation (see the discussion of PP semantic roles below).

[3] A general grammar for English was used. This is slightly different from the existing KANT applications in which controlled input is assumed. However, the architecture is flexible enough to accommodate uncontrolled input as long as the source language grammar can provide the required coverage.

[4] A full discussion of the details is beyond the scope of this paper; the interested reader is referred to the publications available from the KANT home page on the WWW (see final page of this paper for details).

3. The KANT Interlingua

The KANT interlingua is sentential. Each top-level sentence or noun phrase that appears in a text will have one or more interlingua expressions (more than one if the element is ambiguous). An example is shown in Figure 2.

The default rate remained close to zero during this time.

```
(*A-REMAIN                                   ; action rep for 'remain'
   (FORM FINITE)
   (TENSE PAST)
   (MOOD DECLARATIVE)
   (PUNCTUATION PERIOD)
   (IMPERSONAL -)                            ; passive + expletive subject
   (ARGUMENT-CLASS THEME+PREDICATE)          ; predicate argument structure
   (Q-MODIFIER                               ; PP semrole (generic)
      (*K-DURING                             ; PP interlingua
         (POSITION FINAL)                    ; clue for translation
         (OBJECT                             ; PP object semrole
            (*O-TIME                         ; object rep for 'time'
               (UNIT -)
               (NUMBER SINGULAR)
               (REFERENCE DEFINITE)
               (DISTANCE NEAR)
               (PERSON THIRD)))))
   (THEME                                    ; object semrole
      (*O-DEFAULT-RATE                       ; object rep for 'default rate'
         (PERSON THIRD)
         (UNIT -)
         (NUMBER SINGULAR)
         (REFERENCE DEFINITE)))
   (PREDICATE                                ; adjective phrase semrole
      (*P-CLOSE                              ; property rep for 'closer'
         (DEGREE POSITIVE)
         (Q-MODIFIER
            (*K-TO
               (OBJECT
                  (*O-ZERO
                     (UNIT -)
                     (NUMBER SINGULAR)
                     (REFERENCE NO-REFERENCE)
                     (PERSON THIRD))))))))
```

Figure 2: A Sample KANT Interlingua

Each interlingua is essentially a case frame composed of a concept head, features, and semantic roles. The concept is usually lexically specified by the head of the syntactic constituent the interlingua corresponds to. The head is followed by zero or more feature-value pairs or semantic roles. Some features can have binary or atomic values (e.g., FORM, TENSE, MOOD, etc.); such features typically represent grammatical information that is fundamental to the meaning of the utterance. A semantic role is a slot that is filled with an embedded interlingua expression, typically because the meaning of the filler is too rich to be represented as a binary or atomic value. Each semantic role contains an interlingua fragment headed by the concept associated with the head of the syntactic constituent the semantic role corresponds to (e.g., the subject, object and a prepositional phrase in the f-structure can correspond to the `AGENT`, `THEME` and `Q-MODIFIER` slots in the Interlingua). System-specific default values are assumed, e.g., in the Interlingua in Figure 2 the values of the features `PROGRESSIVE` and `PERFECT` are not specified and default to `-`.

The details of the commented example in Figure 2 should be self-explanatory. More details regarding the development of the KANT interlingua can be found in (Leavitt, et al., 1994). In addition, some representation issues with this interlingua and how it applies to the workshop text will be discussed below in Section 5.

It would be fair to characterize the KANT interlingua as a "literal" representation of the input text, which does not attempt to resolve referential phenomena or other types of generalization which require inter-sentential reasoning. Because the current version of KANT is used primarily for dissemination of English texts, the interlingua can be thought of as a frame representation of the English semantics, with a limited amount of canonicalization in the direction of language independence. Of course, this design makes it easy to develop KANT applications for practical MT problems (such as translating technical documentation in a limited domain); but the disadvantage of this interlingua design is that it is underspecified for more general, less technical domains where more reasoning may be required to represent non-literal phenomena.

Despite the literal character, the KANT Interlingua format is particularly well-suited for the representation of the predicate argument structure, the actors in an utterance and the roles they play in the relationship to the action expressed by the main verb (represented as the names of the semantic roles in the Interlingua such as `AGENT`, `THEME`, `BENEFICIARY`, etc). In addition, semantic characteristics of the actors are easily represented in the Interlingua as additional semantic features. These features usually migrate to the Interlingua from the lexical entries from which the concepts appearing in the Interlingua originated. All this is crucial for a successful implementation of selectional restrictions that are often necessary in machine translation in order to be able to select the correct lexical item in the target language. On the other hand, information about temporal relationships, for example, is usually less crucial in a technical domain.

Since the KANT Interlingua has been designed to be used in machine translation, it is often very useful to add features to it that signal some characteristics of the sentence in the input language which are not necessarily semantic or abstract in nature. Figure 2 contains an example of such a feature: `(POSITION FINAL)` is a translation clue that has proved useful in our experience with applying KANT to various domains. It is clearly a syntactic feature, in this case describing the position of the PP modifier in the input sentence. This type of information can be efficiently used to generate the correct position of the corresponding modifier in the target language.

4. Adapting KANT for the Workshop Text

KANT has the following three characteristics which influenced the adaptation of the Analyzer to the text at hand and the resulting representations:

- KANT is designed for well-defined technical domains, such as heavy equipment manuals. The implication is that the vocabulary is limited to a subset of possible words and meanings (e.g., for authoring in a particular domain such as medical records, heavy equipment, etc.). Plus, there is little need for inference (the interlingua can be a "literal" semantic representation, which won't work in domains with more idiomatic/metaphoric/etc. use of language). As a result, lexical coverage of the system has to be ensured before a sentence can be analyzed. In particular, domain-specific phrasal lexical entries and concepts have to be created.
- KANT is designed for dissemination of English texts in multiple languages. The implication is that the interlingua used is in many ways closer to an "English semantics" than a completely language-independent interlingua. The tense information representation is an example. In KANT Interlingua, the tense information is encoded by using the features of `TENSE` and `ASPECT` which represent a view centered on English with not much direct correspondence in other languages. In a more general framework features of this kind might have to be replaced by more general ones describing relationships among a set of time indices.
- KANT is designed for sentence-by-sentence analysis of the input text, and does not represent the larger structure of the document (e.g., paragraph level). As a result, the current version of KANT lacks a representational level which supports resolution of inter-sentential phenomena (such as anaphora, ellipsis, definite descriptions, etc.). While this design decision is justified by the lack of such phenomena in typical KANT domains (Mitamura and Nyberg, 1995; Nyberg and Mitamura, 1996), it does limit the broader applicability of KANT to less restrictive domains (as exemplified by the workshop text).

In order to generate interlinguas for the sentences in the workshop text, we made the following extensions to the KANT analyzer knowledge sources:

- Extensions to the lexicon. We added new compounds, more specific meanings for existing words, and some new verbs and verb alternations (see Section 5.1).
- Extensions to the Interpreter rules, to handle new semantic roles for this domain.
- Manual disambiguation of the text. Interactive disambiguation and SGML tagging (Mitamura and Nyberg, 1995) were used to disambiguate each sentence, so that the single preferred reading was produced.

It should be noted that no procedural changes had to be made. The current results of adapting the KANT analyzer for the workshop text are available on-line (see the final page of this paper for a pointer to the WWW site which contains the interlingua files).

5. Issues

During the process of adapting KANT for the workshop text, several issues arose which are worthy of discussion.

5.1. Domain-Specific Terminology / Concepts

Several domain-specific terms were added as phrases to be represented as atomic concepts, e.g.: *banking client*, *banking system*, *default rate*, *employment opportunity*, *financial service*, *group lending technique*. In general, higher-quality translation is possible when these kinds of terms are represented as single concepts (and therefore treated as single units of meaning in the target language lexicons), with the distinct disadvantage of a higher up-front cost to develop the domain-specific lexicon. Another disadvantage is that this strategy does not migrate well to new domains where there is no customization of the lexicon in advance.

A somewhat related problem is presented by the following sentence:

- *PRODEM used the group lending technique of "solidarity groups" and began making small working capital loans.*

The quoted phrase *"solidarity groups"* represents an example of a new term and concept being introduced in the text. This particular concept had to be added to the lexicon before the sentence could be analyzed using the KANT analyzer. In the Interlingua, the only indication of the unusual character of the phrase is `(QUOTED +)`. This representation is at best schematic, although it might be sufficient for obtaining a correct translation in a language that would use a similar exposition technique (i.e., quoting) for phrases of this kind. However, we do not have a good way of creating such entries on-the-fly, which might be required for less restricted domains.

5.2. Proper Names

Several organization names were added as technical phrases, to be represented as atomic concepts: *ACCION International*, *ACCION*, *Banco Solidario*, *BancoSol*, *Bolivian Superintendency of Banks*, *Superintendency of Banks*, *Calmeadow*, *Fundes*, *InterAmerican Development Bank*, *PRODEM*, *K-Rep*, *Accion Comunitaria del Peru*, *Genesis*.

In some cases, two different names or phrases were used to refer to the same organization (e.g., *ACCION International* vs. *ACCION*, *Banco Solidario* vs. *BancoSol*). The KANT analyzer can represent these as a single concept in the Interlingua only when the two surface forms share the same concept in the lexicon.

For robust coverage of this domain, a complete gazetteer of place names should be added to the KANT analyzer; for this experiment, we simply added by hand those that appeared in the workshop corpus, e.g.: *Latin America*, *Bolivia*, *Canada*, *Switzerland*, *Kenya*, *Guatemala*.

5.3. Structural Ambiguity

There are several places in the text where structurally ambiguous constructions appear, e.g.:

- *... institutions in thirteen countries in Latin America and six cities in the United States*
- *... has lent over $1 billion to microenterprises in the last five years, in loans averaging less than $500*

Resolving these cases automatically requires some reasoning component – e.g., in the first example it is possible to see that the object of the second preposition *in* is *Latin America* (and not the whole conjoined NP) only if it is possible to infer that a country can't be in a city. In cases like these, the correct analysis was selected by hand. As long as the correct syntactic analysis can be chosen, structural ambiguity at the syntactic level does not influence the quality of the Interligua. In contradistinction to some semantic approaches, the KANT Interlingua does not allow for underspecification. In particular, structural ambiguity has to be resolved before the Interlingua is created.

5.4. Prepositional Semantic Roles

One design choice worth noting concerns a shift in the representation of prepositional phrases. In earlier versions of KANT, the system abstracted away from the preposition entirely, representing PPs as semantic roles (such as `LOCATION`, `TIME`, `GOAL`) filled by nominal interlinguas (the contents of the prepositional object). After scaling up the KANT analyzer for use with several target language generators, a change was made whereby PPs are represented as `Q-MODIFIER` semroles, which are filled with special `*K-` concepts representing the entire prepositional phrase; the previous contents of the semrole (the prepositional object) are represented as the filler of the `OBJECT` role in the `*K-` frame.

After significant development of different target languages (French, Spanish, German, Italian, Portuguese) using the abstract representation, it was found that the majority of the prepositional selections were either default, or based on the semantics of the attachment site and the prepositional filler. This lead to a proliferation of semantic roles of the form LOCATED-IN (a location expressed by an "in"-PP-phrase) that essentially carried the same information as the original prepositional concept (e.g., *K-IN) in selecting the target language preposition at the mapping stage. Compounding this issue was the fact that a little over one hundred highly-specific prepositional semantic roles were in use. When generalizing KANT for additional domains, these prepositional semroles were eliminated at the request of the generation developers for the reasons given. It should be noted that this decision is application specific, and also most probably specific to the particular design of the KANT system. In other applications that do not invove the difficult task of generating the correct preposition in the target language, the abstraction of the prepositional meaning might be better justified.

However, in the case of the oblique prepositional phrases that show a strong semantic relationship with the verb, the preposition can be lexically specified and analyzed as the complement of the verb. As a consequence, it is possible to assign a semantic role to the object of the preposition in relation to the verb (e.g., in the representation of *lent $1 billion to the microenterprises* the concept *O-MICROENTERPRISE can be placed in the BENEFICIARY slot of the interlingua headed by *A-LEND).

6. Summary

In the experiment we described, an Interlingua representation was automatically produced for all sentences in the sample text (excluding paragraph titles). The format of the Interlingua is dictated by the needs of an MT system processing documents in a technical domain. The representation is somewhat "literal" in the sense discussed above and some phenomena received only a very simple treatment. According to our experience, the resulting representation is sufficiently abstract and detailed to enable automatic translation into languages like German, French or Spanish in the KANT system when combined with suitable, domain-specific knowledge sources (a generation lexicon, grammar, and mapping rules). As we expected, the main difficulty in applying the KANT analyzer to the sample text was in the lexical coverage of the new domain and in the creative and unrestricted use of language. As for the representation of phenomena like tense or prepositional semantic roles, our experience with the KANT architecture tells us that the English-oriented representation can be sufficient in a restricted domain in a system that performs an additional rearrangement of the information contained in the Interlingua before the TL generation step (mapping into the TL f-structure). However, we realize that the representations we suggest might not be sufficient for other applications involving, e.g., knowledge acquisition from text or qualitative reasoning. The main deficiency we can see is the lack of any resolution of referential phenomena such as temporal ordering of events and deixis. This can also be a problem in an MT setting in which the TL is not "close enough" to the SL. However, we are convinced that the representations we suggest provide a sufficiently strong mechanism to express argument-predicate structure, clausal structure of the input sentences and different kinds of constituent modification.

Also, the Interlingua representations we provide offer an example of output produced by a deployed system that has to deal with constructions that are less interesting from the linguistic point of view but have to be handled in order to enable processing of real texts. The examples of such constructions in the sample text include different partitive constructions (e.g., *three of the most advanced institutions*, *one of the most successful of these institutions*, *40% of all the banking clients*), different types of relative clause modification (*loans averaging less than $500*, *the income generated by this sector*, *the market that needed its services*, *Calmeadow from Canada, which had been very instrumental in the formation of the bank*, *joint venture created in 1986 by prominent members of the Bolivian business community and ACCION International*), expressions involving sums of money and constituent coordination at different levels.

On-Line Resources

- **http://www.lti.cs.cmu.edu/IRW/**
  This site contains the original position paper for the workshop, the interlinguas for the workshop text, and a softcopy version of the final paper (Postscript and PDF).
- **http://www.lti.cs.cmu.edu/Research/Kant**
  This site contains additional information and publications concerning the KANT machine translation system, including several of the papers referenced in the bibliography.

References

Lonsdale, D., A. Franz and J. R. R. Leavitt (1994). "Large-scale Machine Translation: An Interlingua Approach", *Proceedings of IEAAIE-94*.

Leavitt, J. R. R., D. Lonsdale and A. Franz (1994). "A Reasoned Interlingua for Knowledge-Based Machine Translation", *Proceedings of CSCSI-94*.

Mitamura, T. and E. Nyberg (1995). "Controlled English for Knowledge-Based MT: Experience with the KANT System", " *Proceedings of TMI-95*.

Mitamura, T., E. Nyberg and J. Carbonell (1991). "An Efficient Interlingua Translation System for Multi-lingual Document Production," *Proceedings of Machine Translation Summit III*, Washington, DC, July 2-4.

Nyberg, E. and T. Mitamura (1996). "Controlled Language and Knowledge-Based Machine Translation: Principles and Practice", *Proceedings of the First International Workshop on Controlled Language Applications*.

Nyberg, E., T. Mitamura and C. Kamprath (1998). "The KANT Translation System: From R&D to Large-Scale Deployment", *LISA Newsletter*, Vol. 2:1, March.