INTERLINGUA VS TRANSFER? KNOWLEDGE SHARING ACROSS PROJECTS

Nadia Mesli
IAI
Martin-Luther-Straße 14
D-66111 Saarbrücken
Germany
email: nadia@iai.uni-sb.de

Abstract

Sharing knowledge across projects is often considered impossible, especially if the systems involved present differences in underlying theory, representation, or even programming language. This paper, taking a collaboration between the CAT2 and PANGLOSS MT projects as an example, demonstrates that it is possible to combine resources developed separately to support the processing of different modules in a joint system. Despite the differences in the two MT methods, the transfer-based CAT2 and the interlingual PANGLOSS systems proved to share similarities in their linguistic representations that simplified the interfacing between the two systems. Referring to this result the paper presents a discussion of the transfer/interlingua continuum and attempts to bridge the gap between the two MT alternatives.

1 Introduction

As Machine Translation (MT) matures, and as systems become more modular and well specified, pressure increases on system builders to re-use previously built modules instead of developing new ones for every new system. The natural question arises: can one put together a parser for one language and a generator for another to produce an MT system from scratch? What is required of the system? When combining two systems using different MT methods, one may well wonder whether the two systems involved must share some similarities in their linguistic representations, or whether they can offer interfacing facilities that simplify the mapping from one system into the other.

This paper presents the result of a six-month exchange program between the University of Southern California (USA) and the University of the Saarland (Germany). It is an extract of the final report [Mesli 94]. The aim of the project was to test the possibility of combining the CAT2 MT system developed at the Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung (IAI) of the University of the Saarland as a sideline to the EUROTRA MT project with modules of the PANGLOSS MT system developed at the Information Sciences Institute (ISI) of the University of Southern California.

Our purpose was to investigate the possibility of linking together two such different, separately developed systems as the transfer-based CAT2 and the interlingual knowledge-based PANGLOSS

systems, despite differences in underlying method of translation, representation, and even programming language (CAT2 is written in Prolog and PANGLOSS primarily in Lisp). Despite the differences in the two MT systems and methods involved, such a linkup seemed to be possible from the outset, since both systems are modularized, containing several levels for potential interfacing. That CAT2 is transfer-based rather than interlingua-based means that the concept of an interface between source language and target language is inherent in the architecture of the system. Because of the collaboration between different research groups within the PANGLOSS MT project, the PANGLOSS MT system is also designed for having an interface that accepts any kind of information that need to be made available to PENMAN's sentence generator. For this purpose, it uses the normal interface to the Systemic-Functional grammar Nigel called the PENMAN Sentence Plan Language (SPL). Beside the SPL interface notation, PANGLOSS provides other interfacing tools that simplify the mapping of analysis output specifications into PENMAN SPL input specifications.

In addition, the CAT2 interface structure and PENMAN SPL representations are both feature-based functional structures derived from Systemic-Functional grammar descriptions, and the PENMAN generation process is driven by information contained in a knowledge base referring to language-specific as well as domain-dependent knowledge. Therefore, it became clear that, for defining a suitable interface between the CAT2 German analysis module and the PENMAN English text generation system for translating German into English, we had to convert or re-interpret the CAT2 interface feature structure in terms of a knowledge-based environment, using the Sentence Plan Language, which is the normal interface to PENMAN for application programs.

The CAT2-PANGLOSS experiment was carried out for a test sample of 67 German sentences analyzed into 73 CAT2 objects. These sentences were first translated into English by the CAT2 German-English translation module currently under development at IAI in order to guide the translation process between CAT2 outputs and PANGLOSS representations and to compare the results of the two MT methods involved. The grammatical coverage included the treatment of declarative or negative main and subordinate clauses, diathesis (active and passive voice), modality, semantic tense and aspect, lexical control, agreement, determination, pronominalization, and modifier attachment.

2 Architecture of the CAT2 and PANGLOSS MT Systems

2.1 The CAT2 machine translation system

CAT2, a sentence-level and transfer-based multilingual machine translation system, was first developed in 1987 as a sideline to EUROTRA, the massive MT project sponsored by the CEC in Luxemburg encompassing nine languages in twelve countries [Sharp 93]. Like that of EUROTRA and the other EUROTRA sidelines, the architecture of the CAT2 MT system is stratificational. In such a system the translation process is divided into several successive mappings that pass an input through a number of linguistic representations to the final output. This stratificational architecture assumes the transfer-based approach to machine translation, which is opposed to the interlingua-based approach.

The transfer-based approach refers to the fact that the constituent structure (CS) of one language is not an appropriate representation for its translation into another language. For this reason, CS

is transformed during analysis into a more abstract representation, called the Interface Structure (IS), which abstracts away from language-specific structural relations and represents the sentence in a more suitable form for translation. In this structure, each predicate and its arguments and possible modifiers are represented at the same structural level. The IS is essentially a semantically interpreted functional structure that provides information about semantic or thematic roles, noun phrase determination and quantification, sentential time reference, or selectional restrictions.

Both the transfer- and the interlingua-based approaches involve analysis of a source text to a deeper level and synthesis of a target text from that level. The interlingua-based approach, however, uses the interlingua (IL) as a pivot structure that represents a language-independent conceptual structure of the source and target text. Less ambitious than the interlingual approach, the transfer-based paradigm relates the IS of one language directly to the IS of another language via a system of transfer rules. In such a system, the IS is so designed that it can also serve as a logical form or as an interface to a database query system or other Natural Language Processing (NLP) systems. However, the following question arises: is the IS so designed that it can also be an appropriate basis for linkage to an interlingual system's IL?

2.2 The PANGLOSS machine translation system

The PANGLOSS machine translation project is funded by the US Advanced Research Projects Agency (ARPA); it started in late 1991 for an initial three-year period. It is a three-way collaboration between three research groups: the Center for Machine Translation (CMT) at Carnegie Mellon University (CMU), the Computing Research Laboratory (CRL) at New Mexico State University (NMSU), and the Information Sciences Institute (ISI) of the University of Southern California (USC). The PANGLOSS MT system was built initially for translating Spanish into English. Japanese as input language is being added starting October 1993.

The PANGLOSS MT system is an interlingual, i.e., knowledge-based, MT system using knowledge resources of the three participating research groups. This interlingual MT system is characterized as language-neutral, and is designed so that it is possible to translate from a source text to its interlingua text without regard to any eventual target language, and to translate from the interlingua text to a target text without regard to the original source language. The interlingua notation used in PANGLOSS consists of terms defined in a large taxonomy of symbols called the Ontology [Knight 93]. The taxonomization of the topmost regions of the Ontology is linguistically inspired. This means that the location of any term in the taxonomy is determined by its linguistic behaviour in the languages treated, so that the taxonomy partitions the entities of the domain in ways that facilitate their treatment by the system.

USC/ISI's PENMAN system [Penman 89], [Matthiessen & Bateman 91] is currently being used as the output generator of the PANGLOSS MT project. PENMAN is a natural language sentence generation program developed at USC/ISI since 1982. The PENMAN system follows the theory of Systemic Functional Linguistics [Halliday 85]. As input to the text generation process PENMAN takes abstract sentence specifications in the Sentence Plan Language notation. SPL expressions are lists of terms describing the types of entities and their particular features to be expressed in English. The interpretation of SPL terms is performed with respect to the conceptual categories defined in the Ontology. The features of SPL terms are either semantic relations to be expressed,

drawn from the concepts and roles defined in the Ontology, or direct specifications of responses to PENMAN's needs for decision control.

3 Linking the CAT2 and PANGLOSS MT Systems

A major decision had to be made about the way of linking the CAT2 German analysis module developed at IAI and the PENMAN English text generation system developed at ISI for translating German into English. We decided to convert or re-interpret the CAT2 IS representation into the Sentence Plan Language, which is the normal interface to PENMAN for application programs. Thus, the interface would serve as a transfer component between German and English. But how to carry out the translations of the CAT2 analysis outputs in terms of SPL representations, especially if both systems are programmed in different languages? How to translate the specific CAT2 IS features into equivalent PANGLOSS interlingua terms? Does the interface have to handle CAT2 output and PENMAN input specifications in a system-dependent manner that couldn't be used by other language pairs, or does the PANGLOSS MT system already provide mapping tools for all the knowledge modules involved in this project that could be used for our particular interfacing task?

Fortunately, beside the interface notation for PENMAN's application programs, PANGLOSS contains modules that translate system-dependent analysis output representations into PENMAN SPL input specifications. One such module is called the Mapper and consists of two parts: the Lexer and the Parser. Both these tools are written in Lisp. Therefore, in order to achieve our task of linking the CAT2 and PANGLOSS MT systems, we could simply use and modify already existing tools rather than create interfacing tools from scratch.

3.1 The Lexer

The Lexer prepares the analysis outputs for mapping. It reads outputs from various sources — the Spanish Panglyzer for example — and converts them into the canonical factored form expected by the Parser, which is a chart. Because of the system-dependent representations of the analysis outputs, the Lexer program of PANGLOSS cannot be used by arbitrary other modules. Thus, for translating the CAT2 output representations into a suitable form for the mapping, we had to tailor the Lexer program to the formats of CAT2. Given the particular CAT2 IS Prolog notation, the modified Lexer program extracted the result of the IS analysis and converted this Prolog notation into the canonical factored Mapper chart form.

The first step of our translation process was done by producing CAT2 charts from CAT2 IS objects. As an example, the Prolog notation of the CAT2 IS object corresponding to the sentence "Das Buch enthält Ideen" ("The book contains ideas") is converted by the Lexer into the following Lisp chart form, representing all the information contained in each node of the CAT2 tree:

```
(( 0 TEXT (1)  0 0  ((SYN  ((ROLE TEXT)...)))  )
 (  1 S  (2 4 6)  0 0  ((SYN  ((ROLE S)...)))  )
 ( 2 PRED (3) 0 0 ((SYN ((ROLE PRED) (LEX ENTHALTEN)...))) )
 ( 3 LEX "enthalten" 0 0 ((SYN ((ROLE LEX) (LEX ENTHALTEN)))) )
 ( 4 THEME (5) 0 0 ((SYN ((ROLE THEME) (LEX BUCH)...))) )
```

```
( 5 LEX "buch" 0 0  ((SYN  ((ROLE LEX)   (LEX BUCH))))  )
( 6 RANGE  (7)  0 0  ((SYN  ((ROLE RANGE)   (LEX IDEE)...)))  )
( 7 LEX "idee" 0 0  ((SYN ((ROLE LEX)   (LEX IDEE))))  ))
```

3.2 The Parser

The Parser has two functions. First, as a bottom-up chart parser, it reads the charts produced by the Lexer and generates skeletons of corresponding mapping rules. These rewrite rules are written in a unification-based formalism similar to PATR-II and have to be completed by the interface developer for translating the specific analysis output features into equivalent interlingua terms. An example might be the following rules corresponding to the chart of the previous German sentence:

```
((TEXT -> S))
((S -> PRED THEME RANGE))
((PRED -> LEX))
((LEX -> "enthalten"))
((THEME -> LEX))
((LEX -> "buch"))
((RANGE -> LEX))
((LEX -> "idee"))
```

We can see that the mapping grammar consists of rules generated for the text, the sentence, the argument, and the lexical level. The second function of the Parser is then to match these mapping rules with their corresponding input charts in order to produce the final interlingua representations.

3.2.1 The mapping rules

The mapping rules are a particular form of annotated rewrite rule referring to the semantic roles mentioned in each constituent of the CAT2 charts. The task of the interface developer is to complete the skeletal mapping rules in order to translate the specific analysis output features into equivalent interlingua terms. This is done by means of equations putting together the CAT2 attribute/value paths mentioned in each chart constituent and corresponding SPL keyword/value paths.

Excepting the above-mentioned text-level rule (which is not implemented yet in the sentence-based CAT2 MT system and that has therefore not been completed in the interface), the grammar contains mapping rules related to the predicate-argument-adjunct structure of each sentence, to the argument or modifier constituents, and to the lexical entries. The mapping rules related to the predicate-argument-adjunct structure of the parsed sentences translate information at the sentence level. They translate the semantic role system used in the CAT2 MT system (PRED THEME RANGE and their sem fillers, below) into SPL keywords, or more precisely in terms of participant roles of the PANGLOSS Ontology (namely domain and range), and convert all the semantic information specific to sentences into terms of SPL keywords values:

```
((S -> PRED THEME RANGE)
 ((x0 sem) = (xl sem))
 ((x0 sem domain) = (x2 sem))
 ((x0 sem range) = (x3 sem))...)
```

The mapping rules related to constituents translate all the information specific to arguments or modifiers in terms of SPL keywords and Ontology concepts:

```
((PRED -> LEX)
 ((x0 sem) = (xl sem))...)
((THEME -> LEX)
 ((x0 sem) = (xl sem))...)
((RANGE -> LEX)
 ((x0 sem) = (xl sem))...)
```

The mapping rules related to lexical entries mainly translate the CAT2 lexical values ("enthalten", etc.) in terms of Ontology concepts (|contain, bear|). They constitute the bilingual part of the mapping grammar and can be seen as a kind of "transfer" lexical rules:

```
((LEX -> "enthalten")
 ((x0 sem instance) = |contain, bear|))
((LEX -> "buch")
 ((x0 sem instance) = |book, volume|))
((LEX -> "idee")
 ((x0 sem instance) = |idea|))
```

In our experiment, the link between a CAT2 lexical value and an Ontology concept was not achieved by comparing the CAT2 taxonomy with the PANGLOSS Ontology. The choice of the right target concept was rather determined by our knowledge of the appropriate translation of the lexical item (which relied on syntactic, semantic, and contextual criteria) and was achieved by consulting the PANGLOSS Ontology with appropriate tools. This represents a problem for future application of the CAT2-PANGLOSS MT system. Extensions to larger scale require a semi-automatic method of knowledge acquisition to link German lexemes with Ontology concepts, as suggested in [Okumura & Hovy 1994] for linking Japanese lexical entities to Ontology concepts.

3.2.2 The IL representations and PENMAN outputs

Following the generation of the CAT2 charts and the writing of the mapping grammar, the final steps of our translation process between the CAT2 and PANGLOSS MT systems were: (1) to generate the IL/SPL representations corresponding to the CAT2 charts by loading the mapping rules and the charts within the Parser module of the Mapper, and (2) to let PENMAN generate English sentences for these IL/SPL representations. Most of the PENMAN outputs corresponding to our CAT2 charts can be seen as good translations of the German input sentences:

```
PENMAN> (say-spl '(|b-686|  /  |contain, bear|
                   :DOMAIN (|b-687|  /   |book,  volume|
                            :NUMBER SINGULAR
                            :DETERMINER THE)
                   :RANGE (|i-688|  /  |idea|
                           :NUMBER PLURAL
```

```
                        :DETERMINER ZERO)
                    :SPEECHACT ASSERTION
                    :TENSE PRESENT))
"The book contains ideas."
```

The cases of incorrect translations are due to either lexical gaps or incomplete lexical forms in PENMAN, incomplete processing of SPL macro values, grammatical gaps (PENMAN doesn't yet allow several sentence modifiers) or incorrect grammatical processing. In conclusion, we can say that the translation process between the CAT2 and PANGLOSS MT systems was effected successfully.

4 Conclusion

By writing a mapping grammar for a test sample of German sentences, we have shown that linking together two separately developed systems is possible, and indeed with relatively little effort. The success of this collaboration has confirmed the validity of the technique of knowledge sharing used within the PANGLOSS MT project by showing that information from different theories can successfully be mixed to support the processing of different modules in a joint system. In the perspective of an extension of the CAT2-PANGLOSS experiment to a real MT system, future research should include improvements of our experimental mapping grammar to enable the processing of all possible German sentences. The extension of these rules to other kinds of IS representations is possible and easy as long as CAT2 and PENMAN share similar representations of the predicate-argument-adjunct structure. If the argument structure or a particular grammatical representation has to be changed — these cases are characterized as "complex transfer" in transfer-based MT systems — the translation process is still possible, but becomes more complex and less generalizable.

Despite the differences in the two MT methods involved, CAT2 and PANGLOSS proved to share similarities in their linguistic representations that simplified the interfacing between the two systems. As is well known, the transfer process in a MT environment places complex requirements on both the linguistic theories involved and the theories of translation. In the field of machine translation, an interlingua-based system remains an intriguing concept: an intermediary, language-independent representation, an ideal case of interface structure that in theory separates the analytical side of a system from the generative side. Much has been written in research literature about interlinguas, and more specifically about the question whether the interlingua approach is a viable alternative to direct or transfer-based MT systems. Exchanging experience on these two MT alternatives has been the best way for researchers of IAI and ISI to learn that the difference between a system using an interlingua and a system using transfer rules simply seems to be a matter of degree of abstraction. The question is how far the IS of a transfer-based system abstracts away from the surface form of the languages involved and how far it includes general abstract representations such as cognitive categories (spatial, temporal, and causal relations) or pragmatic relations based on the embedding of the text in its linguistic context and the speech act. The more the IS is abstract and includes representations of such general cognitive categories, the easier is the transfer, but the more complex is the system. Therefore the CAT2 system lies somewhere between a shallow transfer-based system and an interlingual paradigm. It integrates in its IS component some general cognitive categories that guide the transfer into target languages. On the other hand

PANGLOSS can be also seen as lying in between a deep semantic transfer-system and a pure interlingual paradigm. Its interlingua is designed in the same way as a transfer-based system by allowing the translation from a source text to its interlingua text without regard to any eventual target language, and by allowing the translation from the interlingua text to a target text without regard to the original source language.

As a result of this work, we have come to view the debates about the difference between the interlingua- and the transfer-based MT approach with a skeptical eye. Despite the terminology used in the various MT theories, the strategies and methods involved are very similar: an interlingua is not as mysterious and unrealistic as related in research literature but contains similarities with other kinds of MT systems that make the interface possible.

Acknowledgements

I would like to thank Hans Haller and Jörg Schütz for having helped me to acquire the DAAD funding of my research program. I am grateful to Eduard Hovy for having simplified my interfacing task by writing the Lexer program and for having provided detailed and helpful comments on the content of this paper. I am also grateful to Richard Whitney for having provided detailed SPL representations of my German input sentences and to Kevin Knight and Masayo Iida for their support in various ways. Remaining weaknesses are solely those of the author.

References

[1] HALLIDAY, M.A.K., 1985: *Introduction to Functional Grammar.* Edward Arnold Press: London.

[2] KNIGHT, K., 1993: Building a large Ontology for Machine Translation. In *Proceedings of the ARPA Human Language Technology Workshop,* Princeton, New Jersey.

[3] MATTHIESSEN, C. & BATEMAN, J., 1991: *Systemic-Functional Linguistics in Language Generation: Penman.*

[4] MESLI, N., 1994: CAT2-PANGLOSS: Towards a Framework for Knowledge-based Machine Translation. To appear in EUROTRA-D Working Papers, IAI-Saarbrücken.

[5] OKUMURA, A., HOVY, E., 1994: Building Japanese-English Dictionary based on Ontology for Machine Translation. In *Proceedings of the ARPA Conference on Human Language Technology,* Princeton, New Jersey.

[6] PENMAN, 1989: *The PENMAN documentation: Primer, User Guide, Reference Manual, and Nigel Manual.* USC/Information Sciences Institute.

[7] SHARP, R., 1993: *CAT2 Reference Manual Version 3.4.* IAI-Saarbrücken.