KBMT-89 - A KNOWLEDGE-BASED MT PROJECT AT CARNEGIE MELLON UNIVERSITY

Sergei NIRENBURG

Center for Machine Translation,
Carnegie Mellon University,
Pittsburgh, Pennsylvania 15213 USA

The KBMT-89 project at Carnegie Mellon University's Center for Machine Translation is devoted to creating a working prototype of a machine translation system with the following specifications:

- Source languages: English and Japanese;
- Target languages: English and Japanese;
- Translation paradigm: Interlingua;
- Computational architecture: A distributed, coarsely parallel system; and
- Subworld (domain) of translation: personal computer installation and maintenance manuals.

The knowledge acquired for the system includes:

- An ontology (domain model) of about 1,500 concepts;
- Analysis lexicons: about 800 lexical units of Japanese and about 900 units of English;
- Generation lexicons: about 800 lexical units of Japanese and about 900 units of English;
- Analysis grammars for English and Japanese;
- Generation grammars for English and Japanese; and
- Specialized syntax ↔ semantics structural mapping rules.

The underlying formalisms that were developed for the use in this system are:

- The knowledge representation system FRAMEKLT;
- A language for representing domain models (a semantic extension of FRAMEKLT);
- Specialized grammar formalisms, based on Lexical-Functional Grammar;
- A specially constructed language for representing text meanings (the interlingua); and
- The languages of analysis and generation lexicon entries, and of the structural mapping rules.

The procedural components of the system include:

- A syntactic parser with a semantic constraint interpreter;
- A semantic mapper for treating additional types of semantic constraints;
- An interactive augmentor for treating residual ambiguities;
- A semantic generator producing syntactic structures of the target language, complete with lexical insertion; and
- A syntactic generator, producing output strings based on the output of the semantic generator.

The support and environment facilities in KBMT-89 include:

- A knowledge acquisition tool for acquiring ontologies and lexicons, ONTOS;

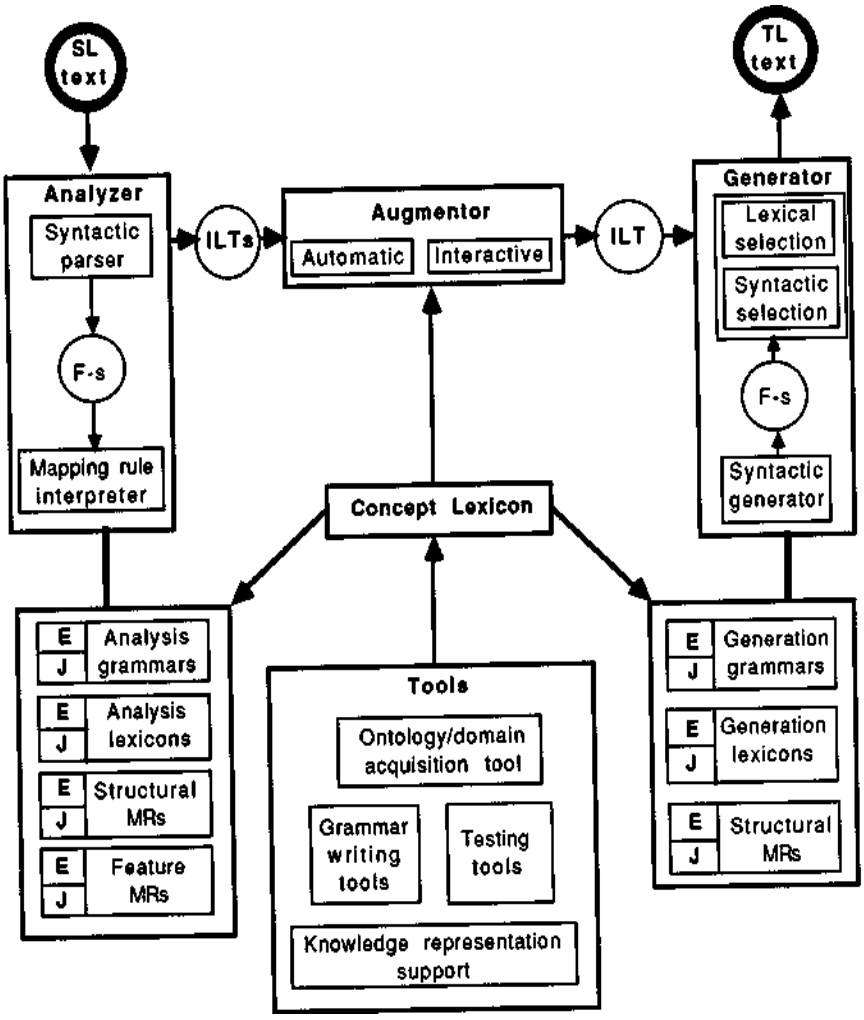

Figure 1: Architecture of the KBMT-89 system. ('SL' and TL' designate 'source language' and 'target language; 'ILT' stands for 'interlingua text'; 'F-s' represents 'f-structure'; 'E' and 'J' designate 'English' and 'Japanese,' the languages used in KBMT-89; and 'MRs' stands for 'mapping rules.'

- A knowledge acquisition tool for acquiring grammars; and
- Testing environments for analysis, augmentation and generation.

KBMT-89 takes as input single sentences of English or Japanese and produces representations of their meanings in a specially devised notation, called *interlingua.* The representation resulting from analyzing a unit of input is called an *interlingua text* or ILT. Taking an ILT as input, the generator produces sentences in Japanese or English that are translations of the original input sentences. Figure
1 illustrates the global architecture of the system.

1. THE ANALYZER

The analyzer consists of two intimately interconnected components — a syntactic parser and a semantic interpreter, called the 'mapping rule interpreter.' The syntactic parser obtains the source language input and produces a syntactic structure for it. The parser uses an LFG-type grammar, so that the resultant syntactic structure is, in fact, an LFG *f-structure.*

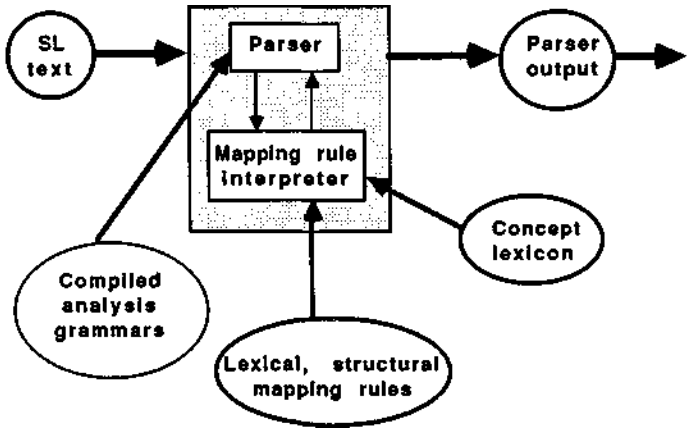

Figure 2: **The architecture of the KBMT-89 analyzer**

As soon as the f-structure for the source language sentence is created, the semantic interpreter starts applying mapping rules in order to substitute source language lexical units and syntactic constructions with their interlingua translations. (This description is simplified for clarity. In reality, mapping rule application starts as soon as an f-structure is produced for any structure component and not after the entire sentence is processed.) Roughly, lexical units map into instances of domain concepts (e.g., the English *data* will map into the interlingua information), while syntactic structures map into conceptual relations (e.g., *subjects* of English sentences often map into the *agent* relations). The process of mapping-rule application is accompanied by elimination of analysis ambiguities through the application of semantic constraints on co-occurrence of various concept instances.

The general architecture of the KBMT-89 analyzer is given in Figure 2.

2. THE KNOWLEDGE PLANE

The meaning of the input text is, as noted above, represented in a specially designed knowledge representation language, an interlingua. In KBMT-89 the interlingua is in turn represented in a frame notation and thus can be viewed as a kind of a semantic network. Like other artificial or formal

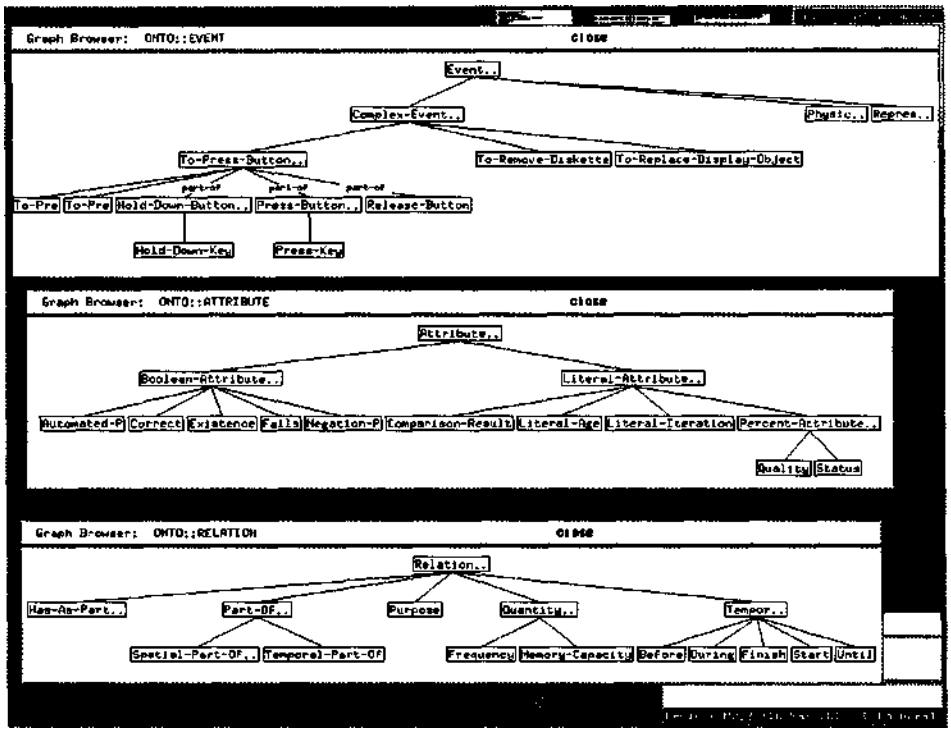

Figure 3: **Representation of a fragment of the KBMT-89 ontology.**

languages, interlingua has its own lexicon and syntax. Yet while the syntax of the interlingua is independently motivated, its lexicon is based on a model of the domain (or 'world') from which the texts to be translated are taken. In the case of KBMT-89 this is the domain of personal computer installation and maintenance. (We sometimes use the terms 'ontology' or 'concept lexicon' to refer to domain models.) Thus, interlingua nouns are *object* concepts in the ontology; interlingua verbs correspond, roughly, to *events* in the ontology; and interlingua adjectives and adverbs are the various *properties* defined in the ontology. The representations of source language inputs, the ILTs, thus contain (numbered) instances of ontological concepts. The ontology itself forms a densely interconnected network of the various types of concepts. Figure 3 illustrates a part of the KBMT-89 ontology. Each of the concept nodes in the figure has, in fact, a much more detailed symbolic representation associated with it.

The syntax of interlingua adds further constraints to the syntactic properties of the general-purpose, frame-oriented knowledge representation language FRAMEKIT (Nyberg [1]) which is used for almost all knowledge representation needs in KBMT-89. The interlingua introduces semantic constraints and marked frame types. Thus, every ILT consists of a *text* frame (In KBMT-89 the inputs were restricted to single sentences, and therefore the need for the text-level index did not arise) and a set of (ILT) *clause* frames. Each *clause* frame has a *proposition* frame associated with it; this, in turn, has a set of *case role* frames attached to it. The heads of the propositions and case roles are instances of ontological concepts, as are many of the proposition and role modifiers. Some of the source language lexical units, however, do not correspond to ontological concepts. Such words can carry special, propositionally relevant meanings (e.g., *be* can be a marker signifying that the following adjective should be understood as a predicative and thus the head of a proposition). They can also carry various

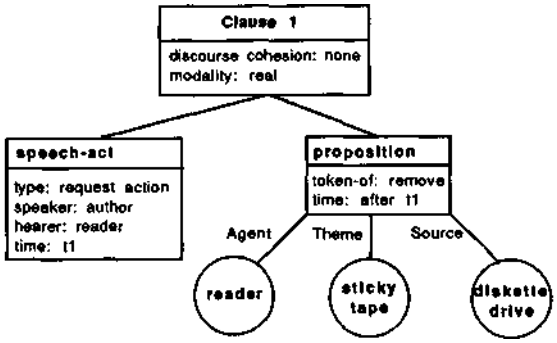

Figure 4: **A sample ILT (schematicized).**

nonpropositional meanings, such as discourse cohesion (e.g., *therefore).* The latter are represented in interlingua using special formalisms not connected with the ontology. An ILT is schematicized in Figure 4.

3. THE GENERATOR

The generation component of KBMT-89 takes an ILT as its input and produces a target language text as its output. Our generator consists of two major modules, one semantic and one syntactic. The former, usually referred to as the 'f-structure builder,' performs the tasks of target language lexical selection and choosing among target language syntactic constructions; it is aided in these tasks by the generation lexicon and the generation structural mapping rules, respectively. The output of this module is an f-structure of the target language sentence that will be output by the system. As its syntactic module KBMT-89 uses GENKIT (Tomita and Nyberg [2]). The KBMT-89 generator is a subset of the DIOGENES generator (Nirenburg et al. [3]).

The architecture of the generation module (which is in many ways similar to the analyzer architecture), is shown in Figure 5; and the process of lexical selection, in Figure 6.

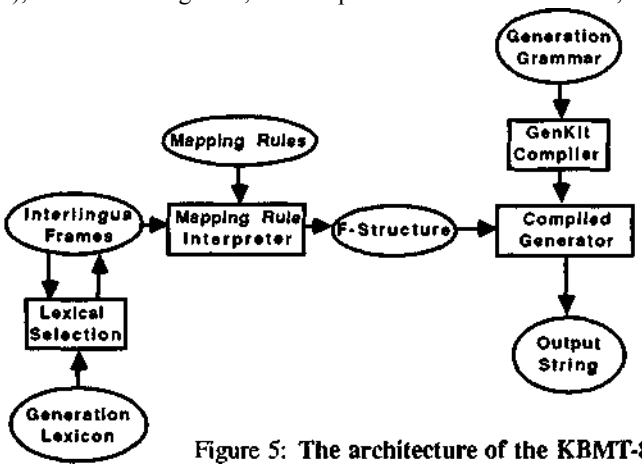

Figure 5: **The architecture of the KBMT-89 generator.**

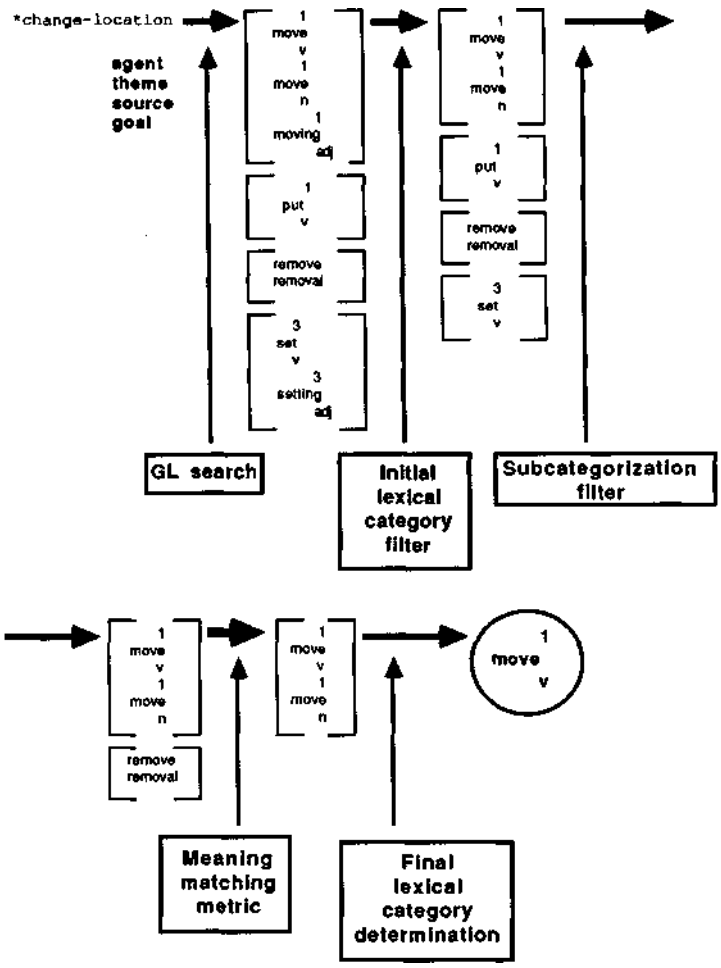

Figure 6: **A representation of the lexical selection process**.

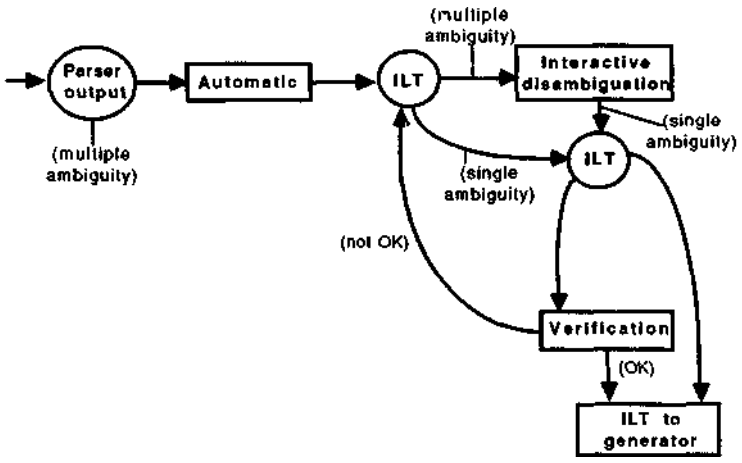

Figure 7: **The architecture of the augmentor.**

4. THE AUGMENTOR

It should perhaps already be apparent that there is a difference between our illustrations of ILT formats and the output from the parser's mapping rule interpreter. There are several reasons for this phenomenon. Among the most important are compatibility between the parser's output structures and the input structures of the generator (that is, the ILTs); constraints on the formulation and applicability of mapping rules in semantic interpretation; and the requirements for representing in interlingua some noncompositional facets of the overall meaning of the sentence, such as speech act and discourse cohesion.

Our augmentor serves two main purposes. First, it reformats the output of the analyzer in the canonical ILT formalism. Second, it helps eliminate residual ambiguities (that is, multiple candidate ILTs for a given input sentence) by applying additional semantic and pragmatic constraints and, if that fails (typically, due to the unavailability of a unit of knowledge), by entering a dialog mode with the users and facilitating their decisions about disambiguation. The architecture of the augmentor is illustrated in Figure 7.

REFERENCES

[1] Nyberg, E. 1988. FRAMEKLT User's Guide. Technical Memo, Center for Machine Translation, Carnegie Mellon University.

[2] Tomita, M, and E. Nyberg. 1988. Generation Kit and Transformation Kit Version 3.2 User's Manual. Technical Memo, Center for Machine Translation, Carnegie Mellon University.

[3] Nirenburg, S., R. McCardell, E. Nyberg, P. Werner, S. Huffman, E. Kenschaft, and I. Nirenburg. 1988. DLOGENES-88. Technical Report, Center for Machine Translation, Carnegie Mellon University.