# Fujitsu Machine Translation System: ATLAS

Hiroshi UCHIDA
*Natural Language Processing Section, Software Laboratory, Fujitsu Laboratories Ltd, Kamikodanaka 1015, Nakahara-ku, Kawasaki-shi 211, Japan*

*Due to the rapid advancement of both computer technology and linguistic theory, machine translation systems are now coming into practical use.*

*Fujitsu has two machine translation systems, ATLAS-I is a syntax-based machine translation system which translates English into Japanese. ATLAS II is a semantic-based system which aims at high quality multilingual translation. In this paper, both the ATLAS-I and ATLAS II translation mechanisms are explained.*

## 1. Introduction

In 1984 Fujitsu marketed the automatic machine translation systems, ATLAS-I and ATLAS II.

ATLAS-I is the world's first commercial English-Japanese translation system, the proto-type of which was completed in 1982. Since then, it has been used experimentally by Fujitsu and selected users. The acquired experience has served to improve the ATLAS system's dictionaries and grammatical rules.

A primary aim in the development phase has been to achieve harmonious interaction between man and machine. For this reason, we rejected the "all or nothing" approach, and adopted a "second best" approach. Here, translation, although rarely perfect is produced swiftly. The text can then be edited up to native speaker quality. This method is both fast and cost-effective.

ATLAS II aims at multilingual translation of three or more languages. At present ATLAS II translates only Japanese-to-English.

This system treats source text analysis and target generation as separate tasks. This feature can be used for translating other languages. Another characteristic is the use of conceptual structure as a bridge between source sentence area and target sentence. Conceptual structure is language independent, and is expressed as a semantic network. When the conceptual structure is too dependent on the source sentence, it is converted to another structure which will be more likely to produce a target sentence. Knowledge, including common sense, is expressed as a semantic relationship between concepts in the same way as conceptual structure.

## 2. ATLAS-I

ATLAS-I uses a syntactic direct approach with some semantic analysis. Its characteristics are:

1) Processing is swift.
2) Grammatical rules can be written easily.
3) The syntax trace of a source text can be reproduced in the translation.

The ATLAS-I translation process is as follows (see Fig. 1): First, an input sentence is segmented into words. At this stage, designated characters such as blanks hyphens and virgules separate the segments.

Next, each word is found in the dictionary, which contains parts of speech, equivalents, and other features. Each word with its corresponding features is concatenated into a string of nodes.

Then, rewriting rules are applied to this node string for part of speech decisions, syntactic analysis, syntactic conversion and equivalent selection. These operations are performed simultaneously from the bottom until all adjacent nodes have been concatenated into one.

Finally, inflections are generated using the inflection table of declinable parts of speech in Japanese; punctuation is generated or deleted, and spaces are inserted.

North-Holland
Future Generations Computer Systems 2 (1986) 95–100

0376-5075/86/$3.50 © 1986, Elsevier Science Publishers B.V. (North-Holland)

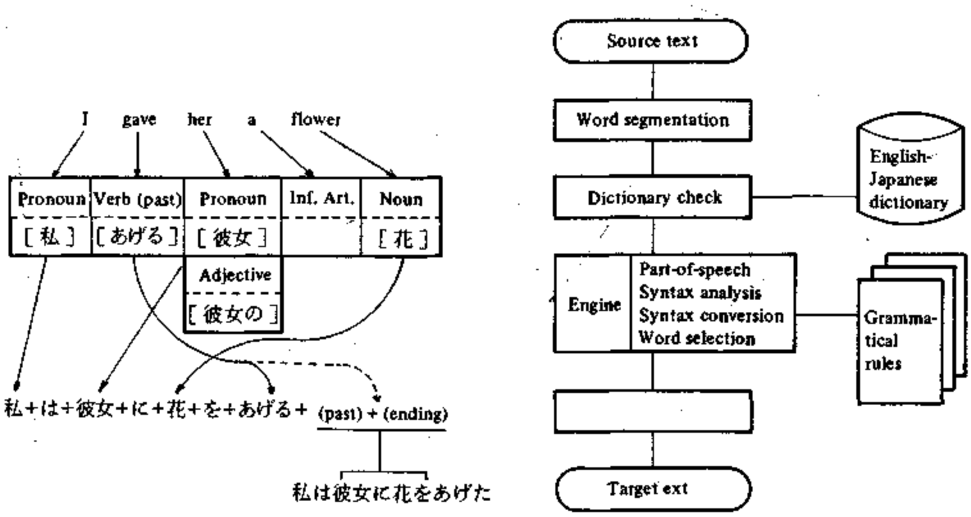


Fig. 1. Translation Process of ATLAS-I.

### 2.1. Rewriting rules

The rewriting rules consist of condition, generation and exit routines.

In the condition field, up to nine node conditions can be specified. Node conditions may be main categories or subcategories. A main category indicates parts of speech and related features (number for a noun; tense for a verb). A subcategory indicates features not classified in the main categories. For example, semantic and syntactic features for a noun, surface-structure pattern for a verb. Both main categories and subcategories specified in the dictionary can be added to or deleted from the rewriting rules.

In the generation field, an arbitrary number of nodes can be specified. Main categories for a node to be generated are defined in this field. Subcategories specified in the condition field can be inherited, or more subcategories can be specified here.

When the conditions for a rewriting rule have been satisfied, the node string specified in the condition field is replaced by a new node string specified in the generation field. Here, words corresponding to each node are rearranged according to the specification in the generation field.

Specific rules control the execution sequence of the exit routine.

### 2.2. Backtracking

When two or more rules satisfy certain conditions at the same time, the rule with the highest priority is executed, and the current node string is reserved. Should execution fail, the reserved node string is recovered from the stack and the rule with the next highest priority is executed. Backtracking can be activated by a rule for a specific condition.

### 2.3. Dictionary

The dictionary is bilingual English-Japanese where each English word is keyed to its Japanese equivalents. The basic dictionary contains about fifty three thousand English words.

Each entry contains several parts of speech; each part of speech has one main category, one standard equivalent word, and an arbitrary number of subcategories. A subcategory indicates syntactic and semantic features of the entry word, and has a subequivalent word which is used for word selection if necessary (see Fig. 1).

In addition to the basic dictionary, a technical dictionary and a user-oriented dictionary are available.

### 2.4. Translation Failure

Translation fails when a string of several nodes cannot be concatenated into one node by applying rewriting rules. In this case, however, an interim result is output based on the final state of the node string. This result can then be edited to produce a correct translation.

## 3. ATLAS II

ATLAS II aims to simulate human translation, understanding a sentence written one language, then expressing it in another. Any natural language is created on the assumption that every person is able to understand a sentence from the context and the meaning of the component words. Language's syntactic regulations are also made on this assumption. To be able to translate naturally a computer should be able to do this.

Humans have their own world models, formed from linguistic knowledge, common sense, cause effect relation, and human characteristics. This is why humans can perform both semantic and contextual analysis with ease. The world model can be extended by inference, or specialized according to the context. Humans also have a language model which guides our use of each word.

ATLAS II is equipped with both a world model and a language model (see Fig. 2).

The world model is expressed as a semantic relation between concepts; the language model as co-occurrence relation between words. Grammatical rules for analysis and generation, and transfer rules are provided for modeling the human translation process.

The conceptual structure is a semantic network representation of an input sentence. Figure 3 shows the conceptual structure which is equivalent to, "I bought a new car". The network consists of nodes and arcs: a node denotes a concept representing the meaning of the words "I", "BUY", "CAR", "NEW", an arc denotes the deep case relation such as ⟨AGENT⟩, ⟨OBJECT⟩, and the junction relation such as ⟨CAUSE⟩, ⟨SEQUENCE⟩. In addition to the above binary arcs there are unary arcs which indicate additional information such as tense, aspect and style. In Fig. 3, ⟨PAST⟩ indicates tense and ⟨ST⟩ indicates focus.

The system understands an input sentence in the form of conceptual structure. Humans understand a sentence by using their knowledge. ATLAS II refers to its world model in the same way as humans. The world model defines every relation between concepts. For example, the knowledge, "Birds fly." is expressed in the world model as follows.

(BIRD, FLY,⟨AGENT⟩) = TRUE

The left-hand side indicates a conceptual structure where an arc ⟨AGENT⟩ conjoins node "BIRD"

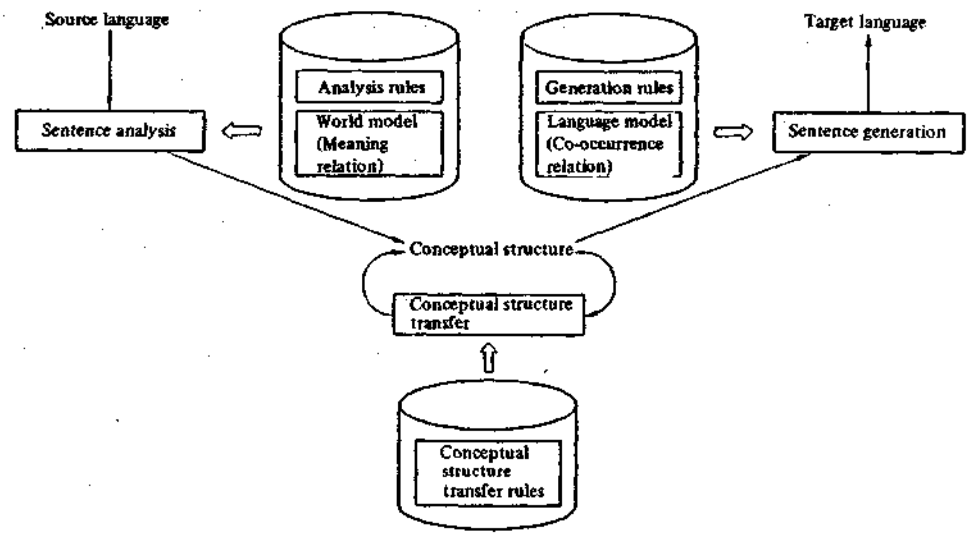


Fig. 2. Translation Process of ATLAS II.

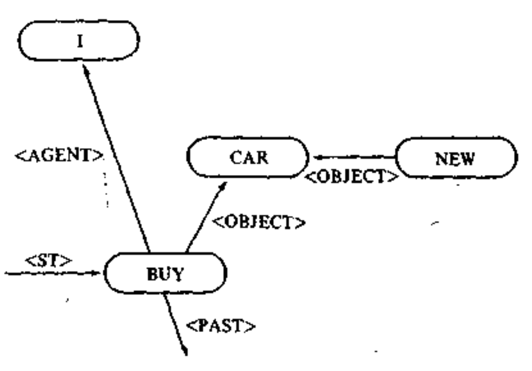


Fig. 3. Conceptual structure.

and node "FLY". This is found to be true by referring to the world model. The system checks whether the conceptual structure is included in the world model. If it is, the system accepts it; if it is not, the system rejects it and asks for another result of sentence analysis.

Relationships between concepts should be as universal as possible. But it is not possible to apply this to all concepts, because each language is to some degree, unique. As a result, a conceptual structure produced by analyzing a Japanese sentence may remain Japanese to some extent; consequently, this structure may not be appropriate for English generation. For example, the sentence "Ningen niwa zunou ga aru." would ideally be translated to "Man has a brain". To do this, conceptual transfer is required; if not, the literal translation "There is a brain in man." will be produced.

Conceptual transfer is performed between conceptual structures: from the source language dependent one to the target language-dependent one. The conceptual structure interface guarantees complete separation between analysis and generation. The pivot approach serves for almost all translations and the transfer approach is used only for specialized ones, allowing a minimum number of transfer rules. As a result, this system is appropriate for multi-language translation.

Figure 4 shows the translation flow of ATLAS II.

### *3.1. Analysis Process*

The sentence analysis section analyzes an input sentence and expresses its meaning as a conceptual structure in the form of semantic network. This section consists of three modules; SEGMENT for morphological analysis; ESPER for syntactic and semantic analysis. This section uses the word dictionary, word adjacency relations, analysis rules, and semantic relations. Figure 4 shows how each module uses the dictionaries and rules and the forms of processing results.

An input sentence is first divided into morphemes. This is a morphological analysis. SEGMENT performs a morphological analysis using the word dictionary and adjacency relations.

Generally, morphological analysis and synthesis are highly language-dependent. This system, however, adopts a language-independent method for multilingual translation. This method uses an adjacency matrix which defines the adjacency possibility between morphemes.

Morphemes extracted by morphological analysis are output in an analysis node list. ESPER receives this node list and each morpheme is treated as a terminal node. The sequence of these nodes is the same as that of the input morphemes. Each node obtains grammatical and semantic information from the word dictionary. Grammatical information is a set of grammatical attributes. This allows each grammatical rule to cover a wide range of linguistic phenomena, thus reducing the number of rules. Each terminal node contains the most probable word of several candidates.

ESPER performs simultaneous syntactic and semantic analysis using analysis rules which are based mainly on context-free grammar. ESPER consists of a status stack, analysis window, and control section. The status stack monitors the status during analysis; the analysis window views two adjacent nodes.

ESPER performs semantic processing and syntactic processing simultaneously. The suitability of syntactic processing is verified semantically.

Semantic processing is performed with a series of semantic symbols which correspond to the conseptual structure. The applied rule attaches a semantic symbol to the new node and determines the semantic relation between two nodes in the analysis window.

The semantic processing checks to find if the processing is consistent with common sense and linguistics.

Finally, ESPER gets the conceptual structure of input sentence. This conceptual structure is veri-

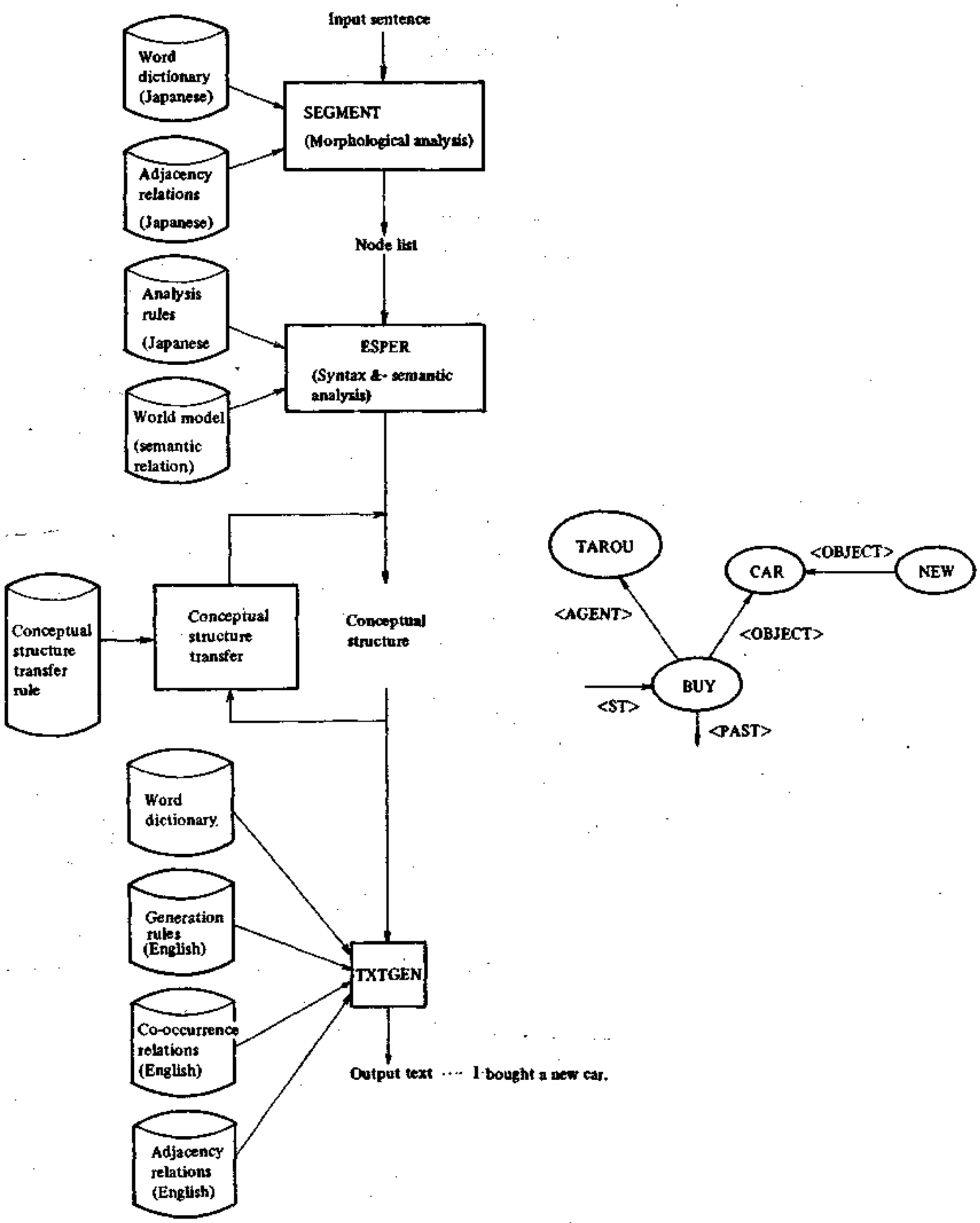


Fig. 4. Translation Flow of ATLAS II.

fied by referring to the world model. If it is incorrect, ESPER reanalyze the sentence and gets another results.

### 3.2. Transfer Process

The transfer section is provided to fill the gap between the source language and the target language. Differences in languages stem from, among other things, the cultural background of the people speaking these languages. Superficially, it appears as a difference in words and grammar; internally, it appears as a difference in concepts and in the speaker's way of thinking.

ATLAS II compares these differences, not superficially, but internally; examining not the differences between words or grammar but the difference between concepts and thinking. The difference, therefore, is treated at the level of the intermediate representation, and the conceptual structure is transfered. However, the pivot approach which does not require this transfer, is suitable for most cases.

We will illustrate some cases which would require such a transfer. For example, the sentence "Heya niwa mado ga futatsu aru." would be literally translated as "There are two windows in this room." But the natural translation would be "This room has two windows".

Another example involves the causative expression. The Japanese language expresses it using the auxiliary verb 'Saseru'; while English depends on an intransitive verb and the word order.

### *3.3. Generation Process*

Target language text is generated from the conceptual structure which is in the form of a semantic network.

This conceptual structure is converted to a linear word string. This direct conversion can eliminate the need for transformations, allowing not only the generation mechanism but also rules to be language-independent.

In this approach, generation rules can deal with both syntactic structuring and morphological synthesizing at the same time, thus simplifying the generation mechanism.

The generation system consists of a generation window, output list and a rule interpreter. The rule interpreter traverses each node of the conceptual structure by moving the generation window and returns with the output list containing the translation results.

The generation window is set at a node of the conceptual structure and is moved from node to node. This window is used to check the nodes and arcs. The output list stores each word in the order of generation. The contents of the output list indicate the surface-structure word order.

The rule interpreter interprets each generation rule, traverses each node by moving the generation window, and selects words from nodes and arcs by checking the co-occurrence relation and adjacency relation. Each selected word is added to the output list.

The coocurrence relation between two words defines the true/false value of whether the two words can cooccur in the same sentence.

Generally a concept includes several words. For example, a concept indicating 'Sonzaisuru' in Japanese includes selection of a word from several candidates by checking the co-occurrence relation between the candidates.

## 4. Conclusion

The biggest problem with any machine translation system is the quality of the translation. Unfortunately, current technology cannot achieve perfect results. We have to provide assistance functions such as pre-editing, and dictionary compilation.

The quality of translation depends on the accuracy of both rules and dictionaries, as well as the amount of information contained in the dictionary. But this presents another problem: the greater the amount of information, the longer the processing time. It is also difficult to guarantee the accuracy of a large of information.

These problems cannot be solved by one company alone. We must ask for assistance from users, especially in the compilation of dictionaries.

We believe, however, that machine translation will eventually prove superior to manual translation in terms of speed and consistency, and will play an important role in international communication.