ATLAS

Hiroshi Uchida

Fujitsu Laboratories Ltd. Kawasaki
1015, Kamikodanaka Nakahara-ku, Kawasaki, 211, JAPAN

1. INTRODUCTION

ATLAS is a machine translation system which aim at high quality multilingual translation. In order to develop a system which deals with various languages with a high degree of precision, analysis and generation mechanisms must be independent of any language, and linguistic knowledge of one language must independent of other languages. Therefore, ATLAS adopts the interlingua approach and using conceptual structure as an interlingua. ATLAS has a language-independent processing mechanisms with a language-independent dictionary structure.

2. ATLAS SYSTEM

In order for humans and computers to understand text written in natural language, it is necessary to know the meaning of words and the meaning within the contexts they are used. An entry in the word dictionary of ATLAS contains the concepts expressed by a word and grammatical characteristics of the word when it expresses a concept. In the word model of ATLAS, the knowledge necessary for understanding the concept is written in a form understandable by the computer, called conceptual structure. The information necessary for understanding the use of words is provided in the form of grammar rules.

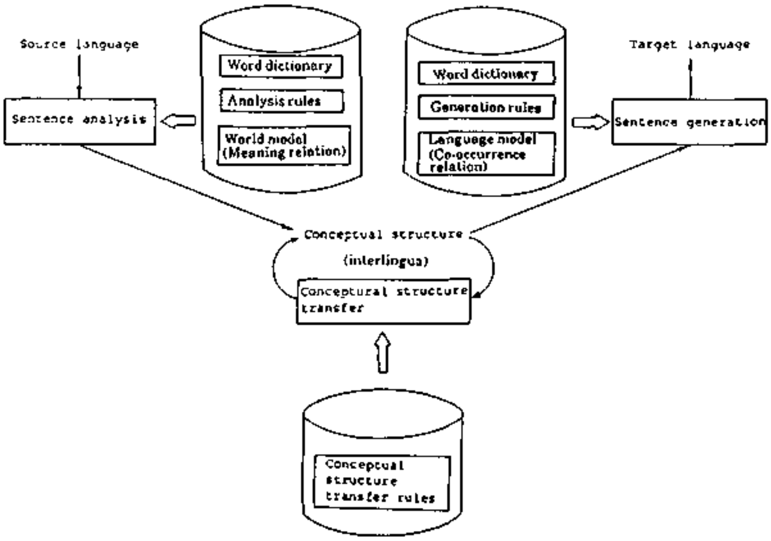

Source language
Sentence analysis
Word dictionary
Analysis rules
World model (Meaning relation)
Word dictionary
Generation rules
Language model (Co-occurrence relation)
Target language
Sentence generation
Conceptual structure (interlingua)
Conceptural structure transfer
Conceptual structure transfer rules

Fig.l The Translation Process of ATLAS

Fig.l shows the translation process of ATLAS. Source language text is analyzed using the word dictionary, analysis rules and world model. The result is expressed as a conceptual structure, which is the interlingua of ATLAS. From the conceptual structure, target language text is generated using the word dictionary, generation rules and language model. If necessary, the conceptual structure is converted to another conceptual structure to fit the target language speaker's way of thinking.

3. INTERLINGUA AND THE WORLD MODEL

The conceptual structure, which is the interlingua of ATLAS, is expressed by a set of binary relations between concepts and features attached to concepts. Fig.2 shows the conceptual structure equivalent to "John drunk beer yesterday." A node denotes a concept representing one of the meaning of words "John", drink", "beer", "yesterday". Arcs denote the deep case relations such as "agent" , "object", and causal relations such as "cause". There are unary arcs which indicate a feature of a concept such as tense and style, etc. In Fig.2, "past" indicates tense and "ST" indicates focus.

In the same way as humans use their knowledge when understanding a sentence, ATLAS refers to its world model when translating a sentence into the interlingua. The world model defines every probable relation between concepts. In other words, the world model contains every conceptual structure for every meaningful sentence. If the conceptual structure of the input sentence is included in the world model, the system accepts it; if it is not, the system rejects it and asks for another sentence analysis.

The vocabulary of the interlingua consists of concepts and relations. Relations between concepts should be as universal as possible. But this universality does not apply to all concepts, because each language has a number of unique concepts. These unique concepts are included as interlingua vocabularies. Some of these unique concepts can be expressed by other concepts or conceptual structure. If the result of analysis contains such a concept, conceptual transfer is performed by using the correspondence between concepts and conceptual structures in the generation process.

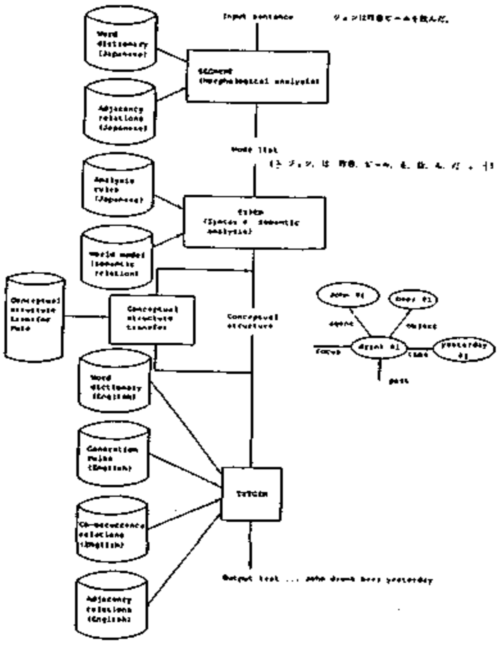

Input sentence
ジョンは昨日ビールを飲んだ。
Word dictionary (Japanese)
SEGMENT (morphological analysis)
Adjacency relations (Japanese)
Node list
Analysis rules (Japanese)
(Syntax & semantic analysis)
World model (semantic relation)
Conceptual structure transfer rule
Conceptual structure transfer
Conceptual structure
john #1
beer #1
agent
object
focus
drink #1
time
yesterday #1
past
Word dictionary (English)
Generation rules (English)
Co-occurrence relations (English)
Adjacency relations (English)
Output text ... John drunk beer yesterday

Fig.2 Translation Flow of ATLAS

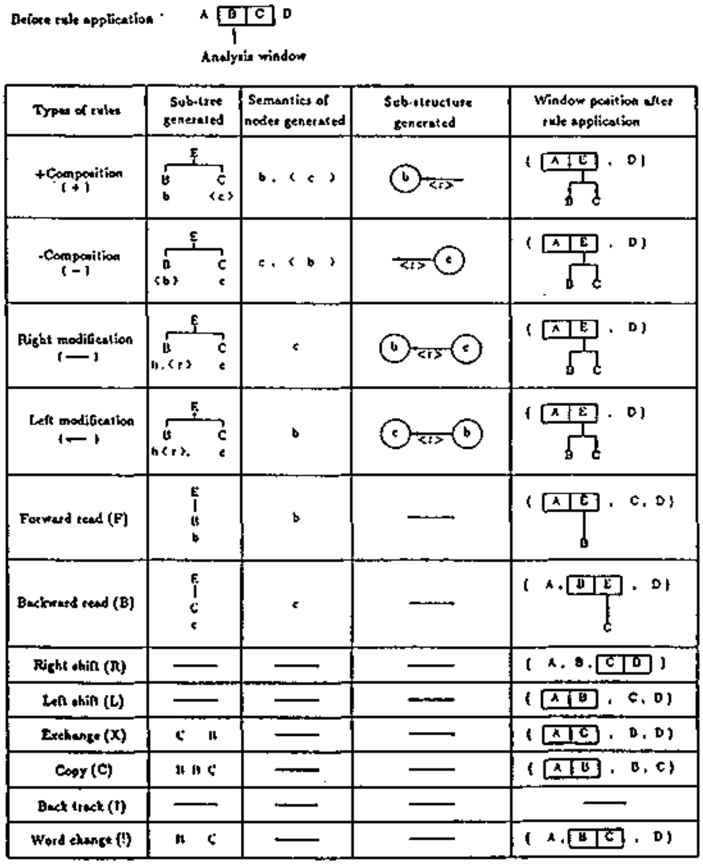
4. SENTENCE ANALYSIS

The sentence analysis phase analyzes an input sentence and produces a representation of its meaning in interlingua. This phase consists of two modules: SEGMENT for morphological analysis; ESPER for syntactic and semantic analysis. Fig.2 shows how each module uses the dictionaries and rules, and the output format.

4.1 Morphological Analysis

An input sentence is first divided into morphemes. SEGMENT performs a morphological analysis using the word dictionary and adjacency relations. Morphological analysis is often thought to be highly language-dependent. This system adopts a language-independent method for multilingual translation.

Staring at the left of the input string, every corresponding morpheme is taken from the word dictionary, and is checked whether it can be adjacent to the leftmost morpheme by referring to the adjacency relations. If it can be, the selected morpheme is removed from the input string and the next matching is performed until no further morphemes are found. Matching is based on the length of the morpheme and the frequency of its appearance. The longest, most frequently appearing morpheme is chosen first. If some strings remain unmatched, the system backtracks to construct an acceptable morpheme list.



Before rule application · A [B C] D

Analysis window

Types of rules	Sub-tree generated	Semantics of nodes generated	Sub-structure generated	Window position after rule application
+Composition (+)	E B C b <c>	b, < c >	b <r>	([A E] , D) B C
-Composition (−)	E B C <b> c	c, < b >	<r> c	([A E] , D) B C
Right modification (——)	E B C b, <r> c	c	b <r> c	([A E] , D) B C
Left modification (←—)	E B C b <r>, c	b	c <r> b	([A E] . D) B C
Forward read (F)	E B b	b	——	([A E] , C, D) B
Backward read (B)	E C c	c	——	(A, [B E] , D) C
Right shift (R)	——	——	——	(A, B, [C D])
Left shift (L)	——	——	——	([A B] , C, D)
Exchange (X)	C B	——	——	([A C] , B, D)
Copy (C)	B B C	——	——	([A B] , B, C)
Back track (↑)	——	——	——	——
Word change (!)	B C	——	——	(A, [B C] , D)

Fig.3 Types of Analysis Rules

4.2 Syntactic and Semantic Analysis

ESPER receives a node list from SEGMENT and performs simultaneous syntactic and semantic analysis using analysis rules 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.

The general format of an analysis rules is :

<CONDITION><GRAM1>+<GRAM2>=<GRAM3><TYPE><RELATION><ACT1ON><PRIORITY>

CONDITION indicates the conditions under which this rule is applied. CONDITION is checked against the message in the basket. GRAM1,GRAM2,GRAM3 are sets of grammatical attributes. GRAM1 and GRAM2 specify grammatical attribute checked against the first and second node in the analysis windows. GRAM3 indicates grammatical attributes for the node created by combining the first and second nodes. TYPE indicates the rule type shown in Fig.3. RELATION shows candidates of modifying relation between the two nodes. ACTION indicates the status after this rule is applied. PRIORITY determines which rule will be applied first when more than one rule can be applied.

ESPER performs syntactic and semantic processing simultaneously. A conceptual sub-structure corresponding to the syntactic sub-tree produced by a rule is generated when the rule is applied. The semantic correctness of syntactic processing is verified by checking whether the conceptual sub-structure is included in the world model or not. When the analysis tree is completed, the entire conceptual structure is again checked against the world model. ESPER backtracks if it is incorrect.

5. SENTENCE GENERATION

The target text is generated from the conceptual structure. Sentence generation is divided into two phases: transfer and generation.

5.1 Transfer Phase

The transfer phase fills the gap between interlingua and the target language. Differences in language stem from cultural background of the people speaking these languages. Superficially, they appear as a difference in words and grammar; internally, they appear as a difference in concepts and in the speaker's way of thinking. If concepts in conceptual structure are not of the target language or the same meaning is expressed by other combination of concepts, the conceptual structure is transferred.

The general format of a transfer rule is:

(Partial Netl, Partial Net2, Relation, Condition)

This rule replaces Partial Netl by Partial Net2 if both Relation and Condition are satisfied.

5.2 Generation Phase

The generation system consists of a generation window to see the node and arcs, output list to stores each word in order of generation and a rule interpreter. The rule interpreter traverses each node of the conceptual structure by moving the generation window and returns the output list of the translation results. Fig4 shows the generation mechanism, which uses generation rules, word dictionary, cooccurrence relations and adjacency relations. In Fig.4, the basket stores message sent from the node itself or from other nodes. The word list is a list of words which express the concept of the node. Both the node and arc names are keys to retrieve words from the word dictionary. Word dictionary entries contain generation symbols which serve as keys to access a generation rule set.

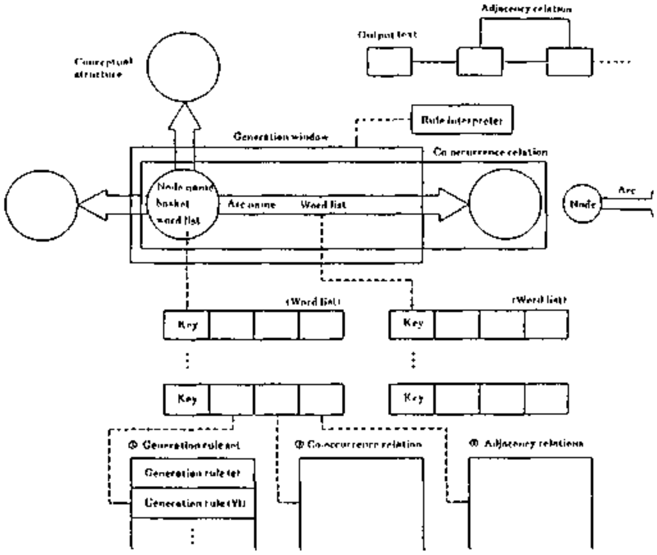


Fig.4 Generation Mechanism

The rule interpreter interprets each generation rule, traverses each node of the network by moving the generation window, and selects words from nodes and arcs by checking the cooccurrence and adjacency relations. The order of traversal is specified by the generation rules. Each word selected is added to the output list.

Cooccurrence relations between two words define the Boolean value of whether the two words can cooccur in the same sentence with a specified relation. In general, a concept may be expressed as several different words. Cooccurrence relations are used to select the most appropriate word. Adjacency relations are used to select appropriate morphemes on the basis of whether two morphemes can be adjacent to each other.

A generation rule set is an ordered set of at least two generation rules. The order specifies the sequence of application, thus determining the word order of the output sentence.

The general format of a generation rule is as follows:

<COND1TION> <ARCNAME> <ACTION> <MESSAGE>

CONDITION indicates the conditions under which this rule is applied. CONDITION is checked against the message in the basket. ARCNAME indicates an arc name to apply the rule. ACTION specifies the type of processing. MESSAGE indicates message to be sent to the basket of the node itself or to nodes connected to the node with arcs.

6. CONCLUSION

We have analyzed and generated text in Japanese, English, French, German, Chinese, Swahili, and Inuit (Eskimo) using ATLAS, with no software modifications. Therefore, we believe that the language-independent mechanism and dictionary structure of ATLAS is suited to multilingual translation.

Translation quality presents the biggest problem to all machine translation systems. Unfortunately, current technology cannot produce perfect results, so post-editing is required. However, post-editing ATLAS translations takes 30-50% less time than full manual translation. Thus, ATLAS is time and cost-effective, even at the current level of technology.