A MACHINE TRANSLATION SYSTEM FROM JAPANESE INTO ENGLISH BASED ON CONCEPTUAL STRUCTURE

Hiroshi Uchida and Kenji Sugiyama

FUJITSU LABORATORIES LTD.
1015, Kamikodanaka, Nakahara-ku
Kawasaki 211, JAPAN

summary

In this paper a language translation system based on conceptual structure is described. The conceptual structure is extended from case grammar from practical viewpoints. The conceptual structure is composed of concepts and relations among them; in our system, a given Japanese text is transformed into conceptual structure, and then an English text is generated from it.

In this paper, the needs and benefits in introducing conceptual structure as intermediate representation are discussed, and then the construct of conceptual structure, and in what way our system utilizes it in a translation process are described.

1. Introduction

It is believed that, in the course of development of present information society, the amount of documents, such as technical writings, correspondences, to be exchanged in every international community has become huge. This great amount of documents have to be translated since no unique universal language is available. But obviously, there is a definite limitation on translation speed by hand. This situation urges on us the importance of development of a machine translation system.

Around 1960 when computers were becoming widely used, various experimental studies were done on machine translation. However, most of them brought about almost no commercial products, and shortly after that, even such kind of researchs seem to have disappeared. Among a few developed systems of that period, only the Russian-English translation systems, MARK2 used by the U.S. Air Force and another used by the Atomic Energy Commission at Aok ridge which was designed at Georgetown University, were widely known. The revised version of the latter system named SYSTRAN, has been on the market, and is currently being used by the EC and other few organizations. At that time, a 'word-for-word' translation was shiefly considered. Such systems are classified in first generation translation systems[2].

After first generation systems, second generation translation systems which rely on intermediate language model instead of a 'word-for-word' translation, are now under development. The features in the approaches of this new generation systems are:

(1) it performs a translation between intermediate languages constructed over source language (SL) and target language (TL) respectively (called transfer approach), as shown in fig. 1,

(2) it encourages to separate linguistic data from programs[2].

In the transfer approach, the intermediate languages are strongly required to retain the characteristics (including syntactical charcterlistics) of the original ones. This approach seems to be effective for translation among languages of the same linguistic family (in the sence that there are similallities in syntax and meaning of a

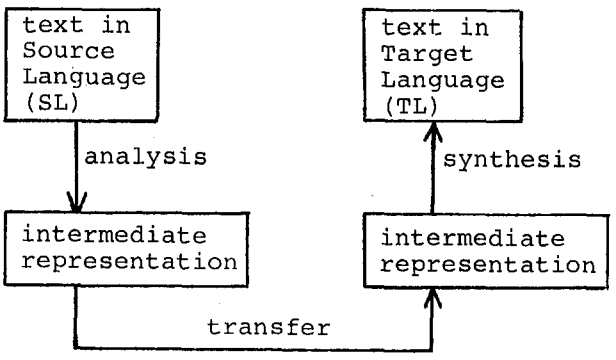

Fig.1 Translation Model of Transfer Approach.

word) such as English, German, and French, because translation of words and some transformation on syntactic structures are merely needed. However, it is not seemed to be quite effective when performing translation among non-related languages, for example, Japanese and English, because of the need for large structural transformation. Among examples of this approach are TAUM of University of Montreal and GETA of Grenoble University.[2]

In our translation system from Japanese into English, conceptual structure is introduced and utilized in translation process.

In this paper, we discuss why conceptural structure is needed, what benifits are obtained from our approach, what conceptual structure is, and in what way our system performs.

2. Need for Conceptual Structure

Let us think of the process we take for natural language translation. Do people really construct intermediate representation for both SL and TL, and then carry out translation between them? Surely they don't. Instead, when translating, they first understand the meaning of a text, and then perform its translation. Furthermore, in addition to explicit meaning in the text, they usually comprehend implicit meaning behind word order (i.e., syntactic information), and a choice of words by the writer. In other words, to understand a text is to extract concepts represented by words or phrases and their mutual relaions from a text.

Therefore, from this obsevation, we can conclude correct translation cannot bypass intermediate semantic representation of sentences (we call conceptual structure) in the process of translation. This conceptual structure is constructed upon concepts in SL, but these concepts are considered as universaly general to the extent that they can be translated into any languages. That is because it is considered peoples will only create concepts which can be understood by every people since they share the same space and physical laws in their life on the globe.

In a case where a concept in one language does not correspond to one in other language, it is supposedly possible to express such a concept with other concepts in another language. In this sence, we can assume every concept has universality so that it is always possible to find a word or a phrase representing a given concept.

As illustrated in fig. 2, in a translation process from language A into B, conceptual structure is constructed from concepts in A, and subsequently is represented with language B. There also exists reverse process, namely from B into A. Some concepts of a language may not fit in any concept of the other one. In this case, it is translated by paraphrasing with available concepts.

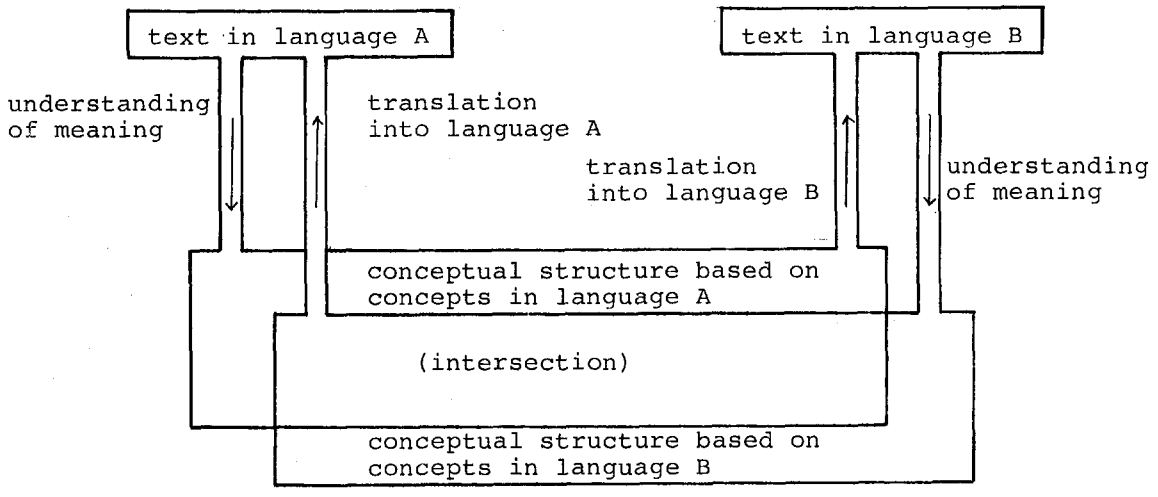

Fig.2 A Model of Human Translation Procedure.

Apparently correct translation cannot be expected only with information of 'surface structure'. Thus conceptual structure plays an important role for translation, but incorporating it into a machine translation system will bring out difficulties, such as how to extract meanings out of sentences, and how to represent meanings in conceptural structure. Nevertheless, we do consider our approach is better than transfer approach, because the latter might involve much more complex treatment, as discussed later. From the above arguments and the fact that our approach is closer to the process we human beings use, we believe ours is more practical and promising than transfer approach.

3. Usefulness of Conceprual Structure

In what follows, advantages in adopting conceptual structure as an intermediate language for a machine translation system are described.

3.1 Separation of Syntax and Semantics

In our approach, first, concepts and relations among them are extracted to construct conceptual structure; next, then it is re-expressed in the target language. In this scheme, syntactical information of the source language does not affect on the second step, and conversely, syntactical regulations of target language do not affect the first step.

However, in transfer approach it tries to convert intermediate structure of the source language to another intermediate structure of the target language, and the translation process cannot be essentially freed from the characteristics of both languages; in other words, syntactical and semantical matters have to be attacked at one time, and this increased complexity seems to make itself inferior to our approach. (However, when SL and TL are in the same linguistic family, transfer approach might be more suitable.)

3.2 Availability of Discourse Information

Discourse information is an inevitable thing in order to comprehend sentences of a natural language, so unless using it, correct translation cannot be expected. In our approach, the meaning of a sentence is represented by conceptural network (which will be defined in section 4), and discourse information will be also composed of these networks.

This scheme, that is, discourse information and sentence meanings are expressed by the same construct, brings us a great convenience to make use of discourse information in process of translation, as well as to accumulate that of sentences transacted so far.

3.3 Advantage in Translation among Many Languages

There are many languages being used in the world, and even if we only count the languages of importance, the number of them is still large. Our translation system aims at translation between Japanese and English language, specifically from Japanese into English, but in the ordinary course of events, it will be applied to other languages in future. When doing translation among many languages the work needed will be beyond our power if transfer approach is chosen, becouse it requires to supply different programs and dictionaries for transfer portion of every pair of intermediate languages.

On the other hand, our approach has only single conceptual structure, so it only requires to add analysis and synthesis procedures for one distinct language (although in our approach, concepts which constitutes conceptural structure might be defined somewhat differently, we believe most of concepts are common). This is one of the advantages over transfer approach in translation among many languages.

4. Conceptual Structure

The meaning of a sentence is expressed by concepts represented by words or phrases, and relations among them. The concept is what is recognized by us abstracting general factors in events and objects (abstraction), but excluding pecurialities to each of them (subtraction).

In our model, a node represents a concept, and an arc represents a relation beween concepts. This constitutes a network representing conceptual structure, and we also call such a network

conceptual structure. This conceptual structure is based on the case grammar by Fillmore[1], but is extended for practical use.

Roughly speaking, concepts in our model is fourfold:

(1) to represent an object (corresponding to a class of nouns),
(2) to represent motion (verbs),
(3) to represent the nature or state of an object (adjectives),
(4) to represent the nature or state of motion (adverbs).

There are relations between concepts of an arbitarily chosen pair of the above classes. For instance, there is a relation between noun and noun class to express "possesion" or "place", and a relation between noun and verb class to express "actor", "place", or "purpose." Some concepts and relations are shown in table 1. The symbols in this table are called semantic symbols.

class	example
verb	give, walk,
adjective	beautiful, red,
noun	I, he, book, Tanaka, Johnson,
adverb	slowly, always,

(a) Concepts (node).
Table 1 Concepts and Relations

class	arc name	explanation
modifier of concepts	past, present, future	shows tense
	temporary, may, must	shows aspect or modal
relation between action concept and others	actor	who does
	causer	who causes an action
	object	the object which is affected by an action
	property	property or state of an action or an object
	to(direction)	direction of action
	from(direction)	
	at(time)	the time when an action occurs
	after(time)	
	before(time)	
	means	method by which a result is obtained
	possesive	possesive relation
	cause	cause of an action
	reason	reason of an action
	modifier	simple modification

(b) Relation between Concepts (arc).

Table 1. Concepts and Relations between Concepts.

4.1 Level of Concept in Conceptual Structure

The meaning of a sentence is expressed by conceptual structure, and deciding in what level such a concept should be expressed is an important matter. That is, the possibility to construct a translation system depends on the level of concepts. For example,

"彼は先生のいうことをよく聞く."

can be expressed with several levels of concepts:

1) let "彼(kare:he)", "先生(sensei: teacher)", and "いうことをよく聞く (iukoto wo yoku kiku: be loyal to, obey)" be concepts,
2) separate "いうことをよく聞く" into "いう(say)", "こと(thing)", "よく(frequently)", and "聞く(listen and obey)".
3) further, separate a verb " " into primitive elements as Schank has done[3] as illustrated in fig. 3, where " " is defined to satisfy one's mind by directing ears toward him.

In the third method, however, although the model is cleared up because of limited primitives, it is not guaranteed that any meaning could be expressed with them. It seems that we could not even know how to choose primitives to express a wide class of meanings. Furthermore, difficulties in sentence synthesis seem to be a barrier for practical use. In addition, when people extract conceptrual structure out of a text in process of translation, they do not seem to separate each concept into elements. From these observations, the above third level of concepts has been rejected for our model.

On the other hand, the oposite direction in terms of level, that is, to introduce compound concepts (usually to represent more complicated concepts) into conceptual structure will make context available from sentences be hidden behind them. Nevertheless, compound concepts are allowed in our system because availability of arbitrary level of concepts enables the system to handle idiomatic expressions and other compound expressions in a straighforward way --- without transforming them into complicated conceptual network (this advantage is also recognized in transfer approach).

As an example, a concept network for

"本システムで使用されているＬＳＩの仕様を表に示す."

is depicted in fig. 4. What this figure explains are:

There is "示す(show)" as in a class of verb whose tense is present, and the place where it occurs is explained by "表(table)"; the object of '示す(show)" is a concept of "仕様(specification)", and it is connected to "LSI" by a 'theme' relation; "LSI" is an object of verbial concept'使用(use)", and has 'aspect' relation of continuation; "システム(system)" has 'modifying' relation of "この(this)".

4.2 Causative Sentence and Voice

Causative sentence is typically recognized by an existence of 'causer' relation.

A sentence is classified into passive or active voice. Sould voice concept be also incorporated in a concept network like tense, aspect, and modal? We consider that writer's choice of voice is not necessarily dominant in conveying meaning; passive voice is often chosen when an actor is of no importance or unnecessary like in "The

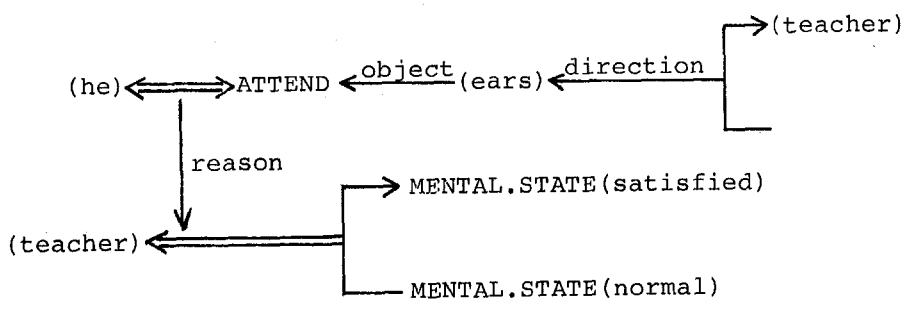

Fig.3 Conceptual Representation with Primitive Concepts.

rocket was launched." At present, so we think the difference of voice is not necessarily explicit in conceptual structure; when generating an English sentence, passive voice is used when actor is omitted in a concept network.

5. Translation Procedure

Our translation procedure is illustrated in fig. 5, and is described in the following. First, the system inputs a Japanese sentence and separates it into 'bunsetsu's, then analyzes relations among them to obtain which 'bunsetsu' modifies which 'bunsets'. This information represents the syntax structure of the sentence, and is output in a form of 'bunsetsu'-table. Based upon this table, concept structure is constructed. Notice, in this structure, syntactical information or words peculiar to the source language is not contained. Next, English phrase for each semantic symbol (attached to a node) is obtained by consulting a dictionary data. In this process, many candidates of English phrase may be found, but the most suitable one is chosen. Further, important grammatical information, such as subject, object, or compliment is set to each arc. Finally, these English phrases are synthesized into an English sentece applying English grammar and modification of words if necessary.

5.1 Analysis of Japanese Language

In Japanese written language, each word in a sentence is not separated by a space like in English; a sentence is usually a succession of words (see 1 of

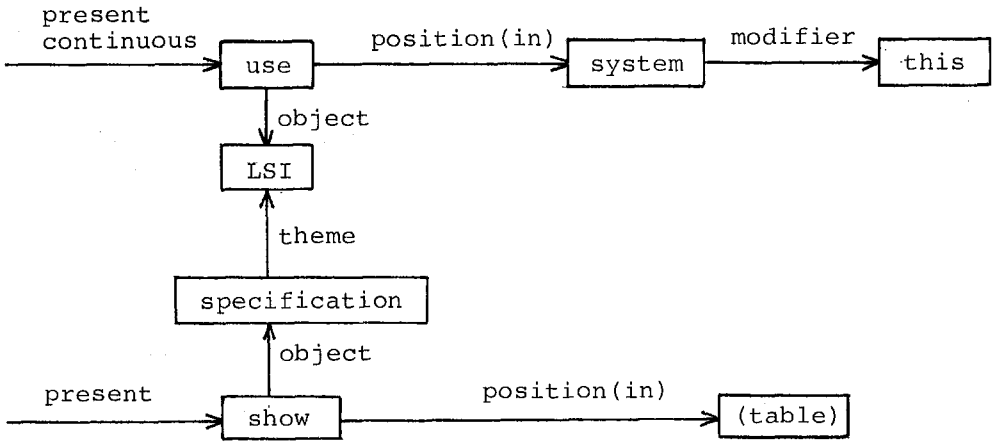

Fig.4 Conceptual Structure.

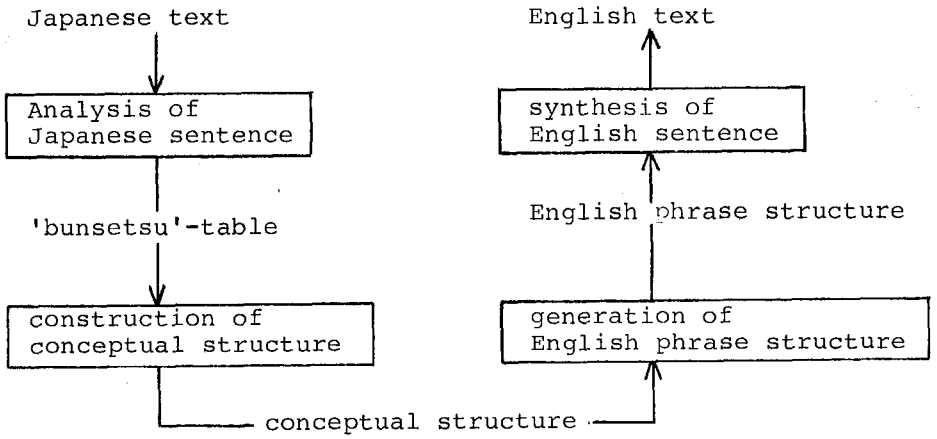

Fig.5 Translation Process.

fig.6). For recognition of each word (see 2 of fig.6), we use adjunctive condition of words; in a Japanese sentence, "私が彼に本をあげた(watashi ga kare ni hon wo ageta: I gave him a book)", "が" can follow "watashi", and "が" can be followed by "彼", but succession of "私" and "彼" is not allowed. This adjunctive relation provides us with very powerful word separation method. (However, since there are many homonyms in Japanese, 100 per cent of correct separation is theoretically impossible. But neary 100 per cent correct separation is being obtained. This matter is not discussed in this paper in more detail.)

'Bunsetsu's are thus recognized as in 3 of fig.6, each of which is composed of 'jiritsu-go' and 'juzoku-go'. Then 'kakariuke'-condition is used to analyze the 'kakariuke' between 'bunsetsu's. As in 4 of fig.6, 'bunsetsu' "私が" does not modify "彼に" nor "本を" but "あげた". This is because 'kakariuke'-condition contains a rule that "私が" only modifies a predicate "あげた" but not others. This 'kakariuke'-condition depends on syntactic features of 'junsetsu'.

In order to identify 'kakariuke' relations more minute information is needed. For example, in fig. 7, "革のカバン (kawa no kaban: a bag of leather)", semantic information should be used to know auxiliary word "の" after "革" specifies the kind of material of "カバン(bag)".

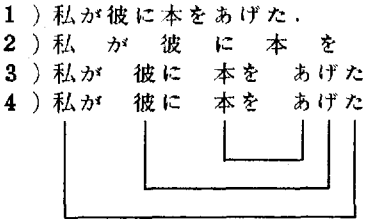

Fig.6 An Example of Japanese Analysis.

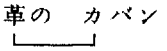

Fig.7 An Example of 'Kakariuke' relation.

5.2 Construction of Conceptual Network

From the 'bunsetsu'-table obtained in the previous step, a condeprual network is constructed with an aid of semantic symbol table which supplies symbols for Japanese words, phrase, or 'kakariuke'-relations.

In a conceptual network, a node represents a concept corresponding to verb, noun, adjective, or adverb of Japanese, and an arc represents functional meaning, such as an auxiliary word " (about)".

5.3 Generation of English Phrase Structure

To generate English phrase structure (i.e., conceptual structure with syntax roles and English phrase attached to nodes and arcs), there is data for each semantic symbol, such as Englsh phrase (possibly a word) and its syntactic type (noun, adjective, verb, and so on). Also, informatin of sentence structure which a phrase takes is provided. That structural information decides the kind of syntax role, such as subject, object, compliment, to be put on an arc. In this phrase structure, verb, adjective, or noun is put on a node, and conjunction, preposition, or relations (such as "which", "where") is put on an arc.

5.4 Synthesis of English Sentence

In accordance to syntactical information given in English phrase structure, English sentence is generated from English phrases put on arcs and nodes. In this process, verb, adjective, adverb, and noun are modified to fit in with a sentence to generate; for example, verb "see" is modified to "saw" if tense is specified so.

6. Conclusion

An experiment of our system on 10 pages text from a computer system manual (approximately 230 sentences included) is currently under way. The results so far is farely good and we would like to comment on this after the data is collected.

One of the possible extension of our system is an automated abstraction system, that is, to generate an abstract on a given text. To do that, we need to

distinguish the equality of concepts of different levels (discussed in section 4) for handling context among sentences. For example, in "There came a girl who was attractive." and "...that honey...", an attractive girl and the honey have to be identified in order to clearify logical relatinship. The conceptual structure thus obtained resembles paragraph structure proposed by Schank[4]. This would be a first step towards an automated machine abstraction of writings.

Acknowledgment

We would like to thank Masato Kobe, and Tatsuya Hayashi who gave helpful suggestions through many discussions, and also would like to thank Sanya Uehara for his assitance in preparing this paper.

References

[1] Fillmore,C.J.:"The case for case" in Bach,E., Harms,R.(eds.): "Universals in Linguistic Theory",Holt, Rinehart and Winston, New York, 1968.

[2] Hutchins,W.J.:"Progress in Documentation Machine Translation and Machine-aided translation", Journal of Documentation, vol.34, No2, June 1978.

[3] Schank,R.C.:"Conceptual Information Processing",North-Holland Publishing Company-Amsterdam,1975.

[4] Schank,R.C.:"The Structure of Episodes in Memory" in Bobrow,D.G. and Collins,A (eds.):"Representation and Understanding",Academic Press, 1975.