COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

A HEURISTIC APPROACH TO ENGLISH-INTO-JAPANESE MACHINE TRANSLATION

Yoshihiko Nitta, Atsushi Okajima, Fumiyuki Yamano, Koichiro Ishihara

Systems Development Laboratory, Hitachi Ltd.
Kawasaki, Kanagawa
Japan

Practical machine translation must be considered from a heuristic point of view rather than from a purely rigid analytical linguistic method. An English-into-Japanese translation system named ATHENE based on a Heuristic Parsing Model (HPM) has been developed. The experiment shows some advantageous points such as simplification of transforming and generating phase, semi-localization of multiple meaning resolution, and extendability for future grammatical refinement. HPM-base parsing process, parsed tree, grammatical data representation, and translation results are also described.

1. INTRODUCTION

Is it true that the recipe to realize a successful machine translation is in precise and rigid language parsing? So far many studies have been done on rigid and detailed natural language parsing, some of which are so powerful as to detect some ungrammatical sentences [1, 2, 3, 4]. Notwithstanding it seems that the detailed parsing is not always connected with practically satisfying machine translations. On the other hand actual human, even foreign language learners, can translate fairly difficult English sentences without going into details of parsing. They only use an elementary grammatical knowledge and dictionaries.

Thus we have paid attention on the heuristic methods of language-learners and have devised a rather non-standard linguistic model named HPM (= Heuristic Parsing Model). Here, "non-standard" implies that sentential constituents in HPM are different from those in widely accepted modern English grammars [5] or in phrase structure grammars [6]. In order to prove the reasonability of HPM, we have developed an English-into-Japanese translation system named ATHENE (= Automatic Translation of Hitachi from English into Nihongo with Editing Support)(cf. Fig. 1).

The essential features of heuristic translation are summarized as in following three points.

(1) To segment an input sentence into new elements named Phrasal Elements (PE) and Clausal Elements (CE),

(2) To assign syntactic roles to PE's and CE's, and restructure the segmented elements into tree-forms by inclusive relation and into list-forms by modifying relation.

(3) To permute the segmented elements, and to assign appropriate Japanese equivalents with necessary case suffixes and postpositions.

The next section presents an overview of HPM, which is followed in Sec. 3 by a rough explication of machine translation process in ATHENE. Sec. 4 discusses the experimental results. Sec. 5 presents concluding remarks and current plans for

enlargements.

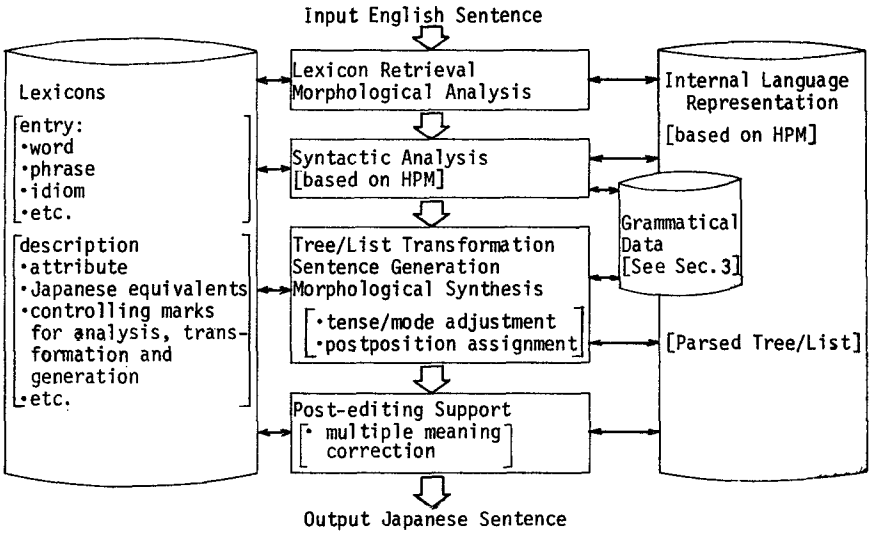

Fig. 1 Configuration of Machine Translation System: ATHENE

2. PARSING MODEL: HPM

To accelerate the clear understanding, an example of the parsed tree on HPM is illustrated in both Fig. 2 and Fig. 3.

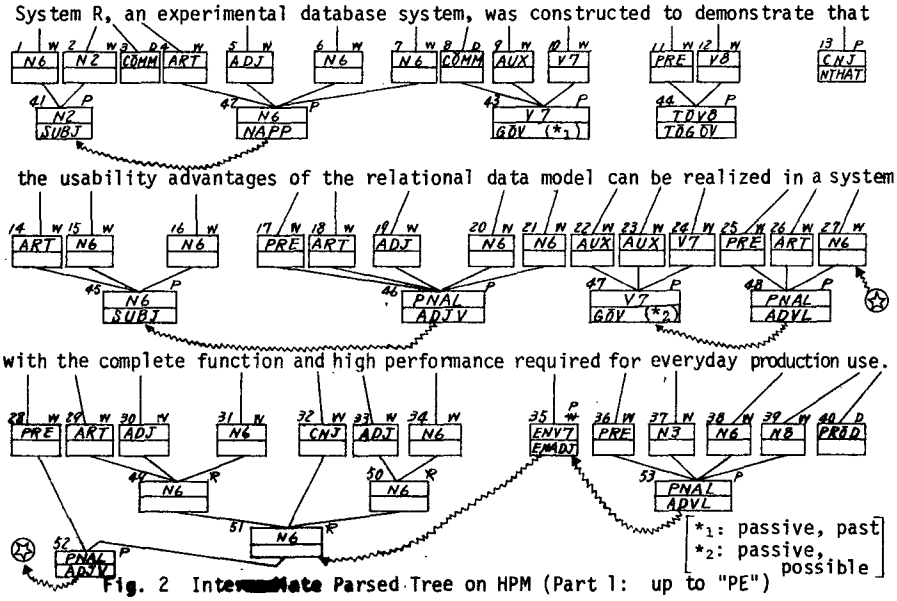

Fig. 2 Intermediate Parsed Tree on HPM (Part 1: up to "PE")

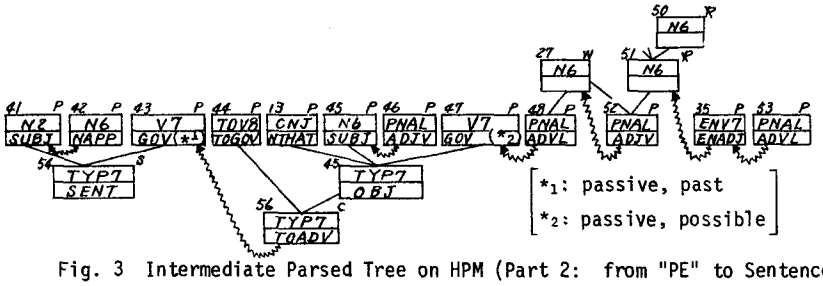

Fig. 3 Intermediate Parsed Tree on HPM (Part 2: from "PE" to Sentence)

2.1 Parsed Tree: A parsed sentence is represented in a "tree" or "list" of nodes linked by pointers. Each node corresponds to a certain "constituent of sentence". "Tree (↓)" is for inclusive relation, and "list (→)" is for modifying relation.

2.2 Constituent: Constituents of sentence is classified into five elements such as: Word Element, Phrasal Element, Clausal Element, Delimiting Element, and Sentence. And these elements have two values: Attribute and Syntactic Role.

2.3 Word Element (WE): WE is the smallest constituent, and therefore is an inseparable element in HPM.

2.4 Phrasal Element (PE): PE is composed of one or more WE('s) which carries a part of sentential meaning in the smallest possible form. PE's are mutually exclusive. Typical examples are: "his very skillful technique (N)", "would not have been doen (V)", and "for everyday production use (PNAL)".

2.5 Clausal Element (CE): CE is composed of one or more PE('s) which carries a part of sentential meaning in a nexus-like form. CE is nearly corresponding to a Japanese simple sentence such as: "~ {wa/ga/wo/no} ~ {suru/dearu} [koto]." CE's allow mutual intersection. Typical examples are the underlined parts in the following: "It is important for you to do so."

2.6 Sentence (SE): SE is composed of one or more CE('s) and is located at the bottom of a parsed tree.

2.7 Dependency Pattern of Verb: Verb-dependency-type code is determined by simplifying Hornby's classification [7], as in Table 1.

Sub-Attr. of V	Dependency Pattern
V1	Be + •••
V6	Vi + To-infinitive
V7	Vt + Object
V8	Vt + that + •••
V14	Vt + Object [+not] + To-infinitive

Table 1. Sub-Attr. and Dependency Pattern of Verb

Sub-Attr. of N	Examples
N1	Place
N2	Person, Organization
N3	Time
N6	Abstract Concept
N8	Means, Method

Table 2. Sub-Attr. of Noun

2.8 Sub-Attribute of Noun: Noun is classified from somewhat semantical viewpoints (cf. Table 2).

2.9 Syntactic Role (SR): SR is important to represent parsing results and to generate Japanese sentences. For example, the sequence of SR such as "SUBJ + GOV + OBJ" will readily imply the Japanese sentence such as "SUBJ + {ga/wa/no} + OBJ + {wo/ni} + GOV". This implication may be quite natural for language-learners.

3. TRANSLATION PROCESS

From the viewpoint of simplicity and maintenability, it might be desirable to describe all the Grammatical Data (GD) in static pattern form. But unfortunately, the pattern form description is lacking in the flexibility to change control structures. Thus we have adopted a combination of "program" and "pattern" to describe GD.

In the followings, we will describe the translation process along with the examples of grammatical data (GD) to be referred. The essential point of the translation process is "to replace some specified node pattern sequences with others, under the appropriate control with grammatical data". This replacement process is composed of following twelve steps:

(1) Text Input: To produce upper-most node sequence in the parsed tree.

(2) Morphological Resolution: To reduce the inflected word to root form.

(3) Lexicon Retrieval and Attribute Assignment: To assign all possible attributes to "WE's".

(4) Ambiguity Resolution in Attributes: To select most likely one from among many possibilities.

(5) Segmentation into "PE's" and Attribute Assignment: To make a PE from matched WE group and give attribute(s).

(6) Re-retrieval of Lexicon: To find again possible WE or PE, especially for "the separated PE" such as "take ∿ into consideration".

(7) Syntactic Role Assignment to PE's: To determine Syntactic Role of PE's by referring a pattern GD as in Fig. 4.

Attr. or Synt. Role Pattern			⟶ Newly Assigned Synt. Role Pattern		
	*			*	
1) N, COMM,	N	, V	⟶ SUBJ, ϕ ,	NAPP	, GOV
2) N ,	PNAL	GOV	⟶ SUBJ ,	ADJV	, GOV
3) V8 ,	{CNJ / that}	, N , V	⟶ GOV ,	NTHAT	, SUBJ, GOV

(* is the Target "PE")

Fig. 4 Pattern to Assign "Syntactic Role" to PE

{Category / Attr. / Synt. Role}	Pattern of PE/CE	⟶ {Attr. / Synt. Role} of CE
1) {PE/CE / ∀ / SUBJ} ,	{PE / V7 / GOV (passive)} , {ϕ}	⟶ {TYP7 / SENT}
2) {PE / ∀ / NTHAT} ,	{PE/CE / ∀ / SUBJ} , {PE / V7 / GOV (passive)} {ϕ}	⟶ {TYP7 / OBJ/COMP}

Fig. 5 Pattern to Make CE with "Syntactic Role"

∀ : anything
ϕ : empty

(8) Segmentation into "CE's" and Synt. Role Assignment: To make a CE from matched PE group and give a Synt. Role by referring patterns as in Fig. 5.

(9) Determination of Modifying Relationships: To determine the appropriate element which the modifier PE should modify.

(10) Construction of Sentence Node (SENT): To complete the whole tree with the root node, SENT.
(11) Tree Transformation: To permute the PE's in each CE. Note that in our HPM, "tree-transformation" is reduced to only a simple repetition of permutation, which has a strong resemblance to language learners' translation methods (Fig. 6).

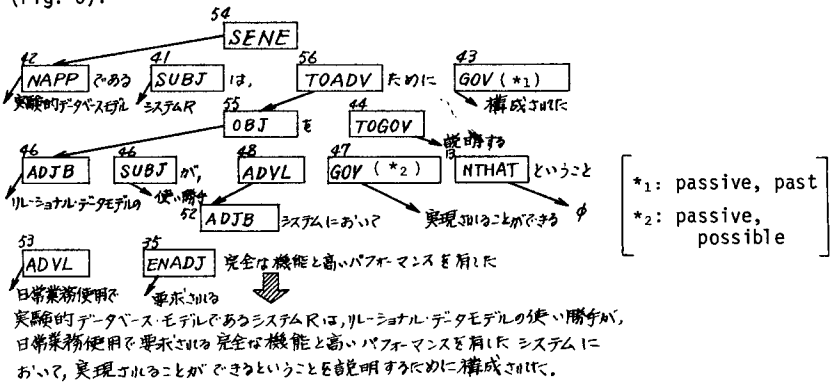

Fig. 6 Tree Transformation and Sentence Generation

(12) Assignment of case suffixes, postpositions and Japanese equivalents.

4. EXPERIMENTAL RESULTS

A prototype machine translation system from English into Japanese named ATHENE, as is sketched in Fig. 1, has been implemented. The lexicons contain nearly ten thousand words, not counting idioms and other multi-word groups, which are mainly composed of educational basic words (up to senior-high-school-level in Japan) and of about a thousand computer terminologies. Our system has translated a series of t t passages extracted randomly from English readers of senior high school and cc puter system journals.

The results of the tests are encouraging on the whole. The system can translate fairly complicated sentences when equiped with the adequate grammatical data and idiomatic phrases. Output sentences, even though far from eloquent style, are worth post-editing, and can be considerably improved with multiple meaning correction through interactive work. Some interesting technical findings are the following:

(1) The following items are sometimes syntactically ambiguous to the system.
  (i) ING + N (ambiguity among ADJ + SUBJ/OBJ, GOV + OBJ, and the like).
  (ii) To-infinitives (ambiguity between adjective and adverbial).
  (iii) Linking scope ambiguity w.r.t. "and", "or", "of" (A and B of C for D).
  (iv) Embedded appositional phrases.
(2) Very long PE's (Phrasal Elements) appear occasionally. (eg. the PE node numbered 52 in Fig. 2 and Fig. 3).

5. CONCLUDING REMARKS

In this paper we try to contend that machine translation should be studied from more heuristic side, or from actual language-learner's methodology side rather than from purely rigid liguistical analysis side. Researchers of very "high level" linguistic analysis side, as is pointed out by Boitet [8], "seem too often to concentrate on interesting high level phenomena as anaphoric reference, discourse

structure, causality and reasoning and to forget at the same time persisting and very frequent lower-level difficulties ... ." This "frequent lower-level difficulty" is the very problem to be solved in practical machine translation, and is actually solved easily by naive foreign language learners only with the help of elementary grammatical knowledge. You had better recall that language-learners must solve the whole even though it is incomplete, on the other hand, pure linguists must solve completely even though it is very limited.

In the light of this contention, we have devised a heuristic parsing model named HPM to accommodate the machine translation to the actual human translation methodologies, and at the same time, on HPM we have constructed a machine translation system named ATHENE. Experimental translation by ATHNE shows the following advantageous points of our heuristic approach.

(1) Contribution to the flexibility, simplicity and maintenability in grammatical description.

(2) Contribution to the simplicity and transparency in transforming phase and generating phase.

One of further problems is to extend the grammatical data heuristically, so as to intensify our machine translation system from learner's level to expert's level. Though our system can translate fairly complex sentences, it still commits learner's level errors when encountering difficulties such as ambiguity of prepositional group modification or of word linking scope for conjunction. Heuristic aspects of semantics are also our current interests of research. Especially the case-grammatical idea [9] seems to be useful to refine our syntactic-role assignment process so as to improve the quality of generated Japanese sentences. A kind of semantic code system (or thesaurus) will also be required to be introduced in our lexicons. Space limitation of this proceeding does not allow us to describe our linguistic model: HPM in detail. We are planning to present the more detailed version of HPM together with later improvement in some appropriate journals.

ACKNOWLEDGMENTS

We would like to thank Prof. Nagao of Kyoto University for his kind and stimulative discussion on various aspects of machine translation. Thanks are also due to Dr. Miura, Dr. Kawasaki, Dr. Mitsumaki and Dr. Mitsumori of SDL Hitachi Ltd. for their constant encouragement to this work.

REFERENCES

[1] Kuno, S. et. al., Mathematical Linguistics and Automatic Translation (Harvard Univ. Report, NSF-8 vol. 1&2, 1962 & NSF-9 vol. 1&2, 1963).
[2] Marcus, M.P., A Theory of Syntactic Recognition for Natural Language (MIT Press, Cambridge, MA, 1980).
[3] Sager, N., Natural Language Information Processing (Addison Wesley, Reading, MA, 1981).
[4] Robinson, J.J., DIAGRAM: A Grammar for Dialogues, Comm. ACM 25, 1 (1982) 27-47.
[5] Quirk et. al., A Grammar of Contemporary English (Longman, London; Seminar Press, New York, 1972).
[6] Chomsky, N., Aspects of the Theory of Syntax (MIT Press, Cambridge, MA, 1965).
[7] Hornby, A.S., Guide to Patterns and Usage in English, second edition (Oxford University Press, London, 1975).
[8] Boitet, C. et. al., Present and Future Paradigms in the Automatized Translation of Natural Languages, COLING-80, Tokyo (1980) 430-436.
[9] Fillmore, C.J., The Case for Case, in: Bach and Harms (eds.), Universals in Linguistic Theory (Holt, Rinehart and Winston, New York, 1968) 1-90.