A Pattern-Learning Based, Hybrid Model for the Syntactic Analysis of Structural Relationships among Japanese Clauses

Akitoshi OKUMURA Kazunori MURAKI Kiyoshi YAMABANA

NEC Corp. C & C Information Technology Research Laboratories
4-1-1 Miyazaki, Miyamae-ku, Kawasaki 216, JAPAN
okumura%mtl.cl.nec.co.jp@sj.nec.com

Abstract

This paper presents a model for analyzing Japanese compound sentences by taking advantage of user's examples. Syntactic analysis of the Japanese compound sentences is one of the most difficult problems for machine translation (MT) systems. One particularly difficult problem is selecting, from all possible candidates, the correct global structure of an individual sentence, i.e. recognizing the set of structural relationships actually linking the various clauses in a compound sentence. MT systems are generally equipped with correction tools for users to modify the incorrect results of the system analyses. Users must use such tools over and over again to modify the same kind of sentences, however, because there is no effective method to memorize the previously corrected examples. The authors here propose a pattern-learning based, hybrid model for analyzing structural relationships among Japanese clauses. The model consists of rule-based modules and a learning module which memorizes correct global structures as well as incorrect ones. All structures are memorized in the form of five-region patterns which are represented by salient features. The model is investigated in comparison with the properties of connectionist approaches, and then the validity of the model is supported by the the results of preliminary experiments.

1 Introduction

This paper presents a model for analyzing Japanese compound sentences by taking advantage of user's examples.

Syntactic analysis of Japanese compound sentences is one of the most difficult problems for machine translation (MT) systems. It requires two kinds of analyses: one for local structures, that is, to select, from all possible candidates, the actual set of structural relationships linking the phrases and the predicate that compose an individual clause, and one for global structures, that is, to select, from all possible candidates, the correct set of structural relationships actually linking the various clauses in a compound sentence. Of these two, global analysis is by far the more difficult.

In order to select an adequate global structures, there are several theoretical and heuristic rules by a rule-based approach[1, 2, 3]. Though they are effective for disambiguating some inadequate structures, they cannot always select the most adequate structure. It is difficult to describe all the rules systematically for the structure determination , because the global structures tend to depend on properties of the examples, like text styles and writers' inclinations.

To solve the problems difficult for the rule-based approaches, some corpus-based approaches are proposed, like *statistical MT, example-based MT* and *connectionist* MT for MT approaches[4, 5, 6, 7]. Corpus-based systems are equipped with an inference mechanism supported by some corpora and data. They are reported to give better results for some fields than rule-based systems [8, 9, 10]. On the contrary, the rule-based systems are more effective in the fields where the information necessary for the solution can be described in the dictionary and the rules, like a local structure analysis [11]. Therefore, the corpus-based approaches and the rule-based approaches should be well integrated to make the best use of each advantage [12]. However, it is not yet proposed how to integrate both approaches for analyzing a Japanese compound sentence. Indeed, MT systems are practically equipped with some correction tools for users to modify the results analyzed by the systems. However, the users must use the tools over and over again to modify the same kind of sentences, because there is no effective method to memorize the examples corrected by the users.

The authors here propose a pattern-learning based, hybrid model for analyzing structural relationships among Japanese clauses. The model consists of rule-based modules and a learning module which memorizes correct global structures as well as incorrect ones. All structures are memorized in the form of five-region patterns which are represented by salient features.

First, a rule-based approach for analyzing a global structure of a Japanese compound sentence is introduced as well as its problems. Second, some problems for memorizing the structures from the examples are mentioned. Third, the hybrid analysis model is proposed as well as the MT system configuration. Last, the model is investigated in comparison with the properties of connectionist approaches, and then the validity of the model is supported by the the results of preliminary experiments.

2 A Rule-Based Approach for Compound Sentences

There are two kinds of structural relationships in a Japanese compound sentence: a global structure, that is, structural relationships among the clauses consisting of the compound sentence, and a local structure, that is, structural relationships among phrases and the predicate in the clause. A Japanese compound sentence consists of several clauses. Each clause modifies a right-located clause, according to the non-crossing principle: individual modification arcs don't cross each other. The main clause, including the head predicate of the sentence, is located at the last of the sentence. All clauses include the predicates. The predicates are located at the end of the clauses, while they precede their modifying nominal phrases in the embedded clauses. Figure.l shows a block diagram of a Japanese compound sentence structure. An arc indicates a global relationship.

Grammatically, innumerable clauses can be included in one compound sentence. Structural ambiguities increase according to the number of the clauses. The total number of the syntactic structure ambiguities is the number of global structure ambiguities multiplied by the number of the local structure ambiguities. Practically, there are some sentences including more than five clauses in technical manuals. Therefore, it is an important problem to disambiguate some inadequate global structures, when analyzing a Japanese compound sentence structure.

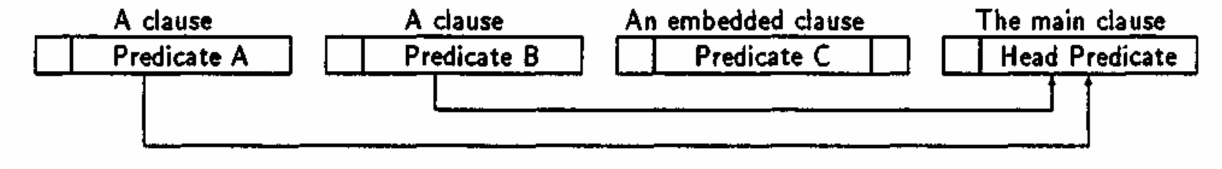


Figure 1: A compound sentence global structure

In Japanese sentences, some specific words help readers to understand the sentences by controlling the relationships among the clauses. They are used as key information for analyzing the global structures[1, 2, 3]. In *Lexical Discourse* Grammar, they are called "functional features". They are represented as predicate inflection forms, auxiliary verbs, conjunctive particles, and so on[1, 2]. They are components of the predicate phrases.

There is a logical preference of attachment (relationship) between a modifier and a modifiee[13]. In Japanese sentences, the logical preference between clauses can be induced by exploiting the functional features[1, 2]. For example, the following sentences consist of three clauses. Although each has two global structure alternatives, correct structures are induced from the properties of "NODE"and "TE" ,as shown in Figures 2 and 3. Example:

1. Ame ga futtaNODE michi ga nureTE hikatteita.
   (AS it rained, the road was wet AND shining.)

2. Ame ga futTE michi ga nureteitaNODE watashi ha koronda.
   (AS it rained AND the road was wet, I fell down.)

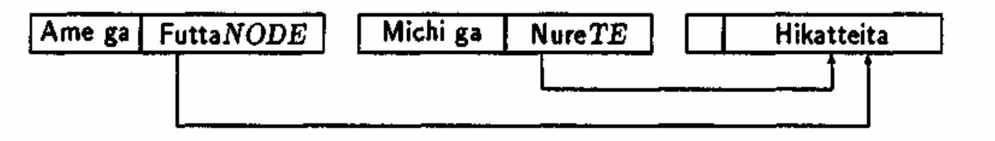


Figure 2: Example 1 syntactic structure

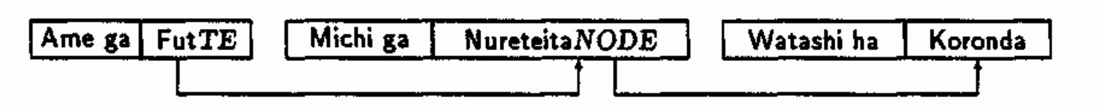


Figure 3: Example 2 syntactic structure

However, the relationship preference is defined as a binary relationship between two clauses; a modifier and a modifiee. It doesn't suggest the most suitable global structure, considering the entire structure of all clauses included in the sentence. Indeed, in order to determine whether one clause syntactically modifies the other, the following expressions included in other clauses should be considered as well as the functional features in the modifier and modifiee clauses.

1. A topic marker, "ha"
2. A comma
3. An ellipsis of a case element, like a subject and an object
4. A demonstrative pronoun and a pronoun

They are control features, which can control the whole relationships. Indeed, the features are partially used for some heuristic rules[3]. It it possible to write heuristic and exceptional rules by the features for an individual sentence. However, it is difficult to describe systematic rules with the functional features and the control features, because there are a lot of combinational variation to be considered.

3 Problems for Memorizing Structures

In order to select an appropriate global structure based on users' examples, there are the following problems for representing the structures. Some of them are related with the problems of *knowledge-based MT* system[14, 15, 16, 17].

1. Flexible representation

   A flexible representation method is necessary to memorize the users' examples, because there are several kinds of compound sentences: some consist of three clauses, and others four or five ... clauses. The memorized relationships of three-clauses sentence should be useful for the structure selection in the four-clauses sentences including the same pattern of the three clauses.

2. Effective representation

   It is necessary to discover a valid set of salient features, which is able to represent the clause properties effectively. Though detailed and direct representation like words and phrases can memorize the clauses very exactly, the representation cannot give any effective result for the sentences which are even a little different from the memorized structures. On the other hand, abstract representations cannot recollect an adequate structure from the memorized structures.

3. Well-grounded representation

   The features should be reasonable enough to be grounded on some theories or persuasive empirical intuition. Otherwise, they don't guarantee any validity for other data, even if they are proved valid for some data.

4. Practical representation

   The features should be extracted from the input sentences automatically and correctly from the practical view-point.

5. Consistent representation with the rule-based modules

   An result recollected from the memorized examples should be consistent with the results analyzed by rule-based modules.

4 A Pattern-Learning Based Hybrid Analysis Model

The pattern-learning based hybrid analysis model consists of the morphological analysis module, the learning module, and the syntactic-semantic analysis module, as shown in Figure.4.

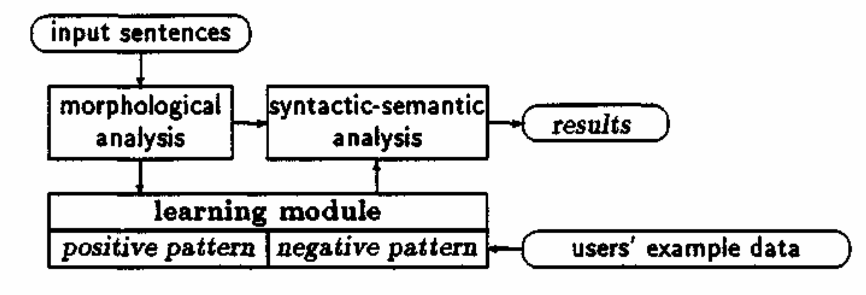
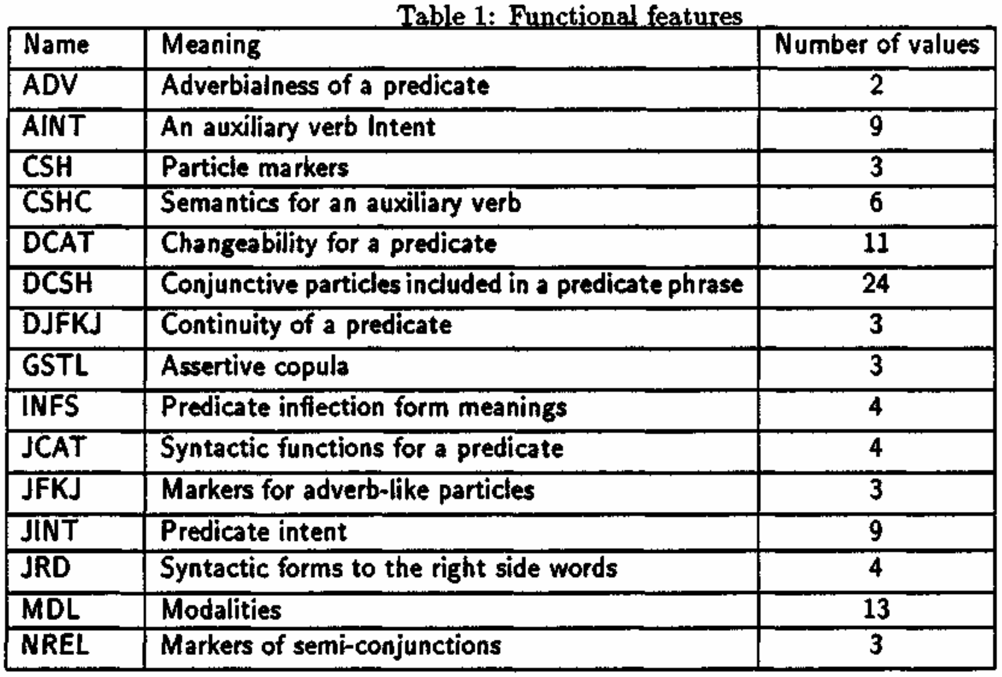


Figure 4: A pattern-learning based hybrid analysis model

Table 1: Functional features



Name	Meaning	Number of values
ADV	Adverbialness of a predicate	2
AINT	An auxiliary verb Intent	9
CSH	Particle markers	3
CSHC	Semantics for an auxiliary verb	6
DCAT	Changeability for a predicate	11
DCSH	Conjunctive particles included in a predicate phrase	24
DJFKJ	Continuity of a predicate	3
GSTL	Assertive copula	3
INFS	Predicate inflection form meanings	4
JCAT	Syntactic functions for a predicate	4
JFKJ	Markers for adverb-like particles	3
JINT	Predicate intent	9
JRD	Syntactic forms to the right side words	4
MDL	Modalities	13
NREL	Markers of semi-conjunctions	3

The morphological analysis module extracts two kinds of salient features: functional features and control features as shown in Tables 1 and 2 from the input sentences.

The learning module memorizes the global structures, based on the users' example data which are morphological analysis results of the users example sentences. The structures are represented by some relationships among two clauses: a modifier clause and a modifiee clause, as shown in Figures 1,2 and 3. Each relationship is memorized in the form of a fixed pattern consisting of five regions; a context-determinant region in front of the modifier clause (Pre-clause region), a modifier clause region, a context-determinant region between the modifier clause and the modifiee clause (Inter-clause region), a modifiee clause region, and a context-determinant region behind the modifier clause (Post-clause region), as shown in Figure 5. All the regions are represented by a series of binary values corresponding to each value of the salient features extracted by the morphological analysis module, as shown in Figure 6. The module independently memorizes two kinds of patterns: "positive" patterns representing structures accepted as correct by users, and "negative" patterns representing the structures which had to be corrected by users.

The syntactic-semantic analysis module inquires of the learning module about the the plau-

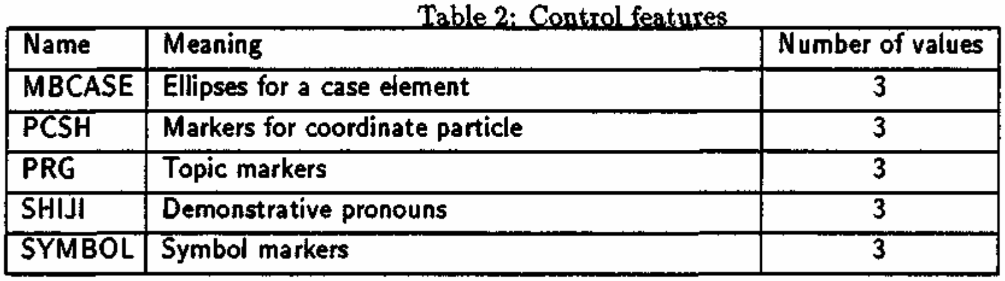


Table 2: Control features

Name	Meaning	Number of values
MBCASE	Ellipses for a case element	3
PCSH	Markers for coordinate particle	3
PRG	Topic markers	3
SHIJI	Demonstrative pronouns	3
SYMBOL	Symbol markers	3

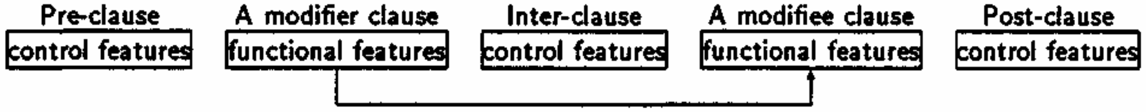


Figure 5: A pattern representation by the features

sibility of relationships, when the analysis module is unable to determine a global structure, all relationships among the clauses. The learning module transmits a positive or negative result about the ambiguous relationships to the analysis module, by referring to the memorized patterns.

The learning module can be directly incorporated into the MT system, as shown in Figure.7. When some ambiguous relationships are remaining unselected by the syntactic-semantic analysis, the learning module works. Whenever the users correct the results of the analyzed global structures, the learning module memorizes positive patterns from the corrected results and negative patterns from the results before the correction. The syntactic-semantic analysis module determines the whole structures: local structures and global structures. After the results are transferred to the conceptual analysis module, the module creates interlingua. Sentences are generated from the interlingua [19, 20, 21].

There are the following characteristics in this model.

1. A hybrid analysis model

   The model consists of the rule-based modules and the learning module. The rule-based modules are parts of general MT systems, which makes it easy to integrate this model into the MT systems. The learning module memorizes the global structures which can not be selected because of the lacks of the rules. The structures are memorized by the the same features that are used in the rules. The results of the module are consistent with those of the rule modules, because the module supplements the lacks of the rules founded on the rule modules.

2. A positive and negative pattern learning

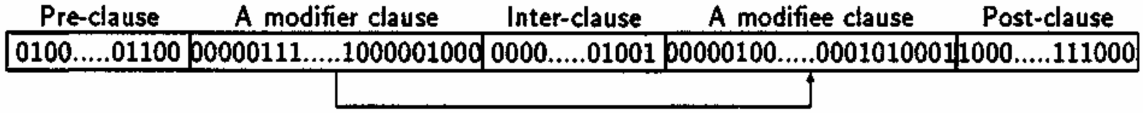


Figure 6: A pattern representation by the binary values

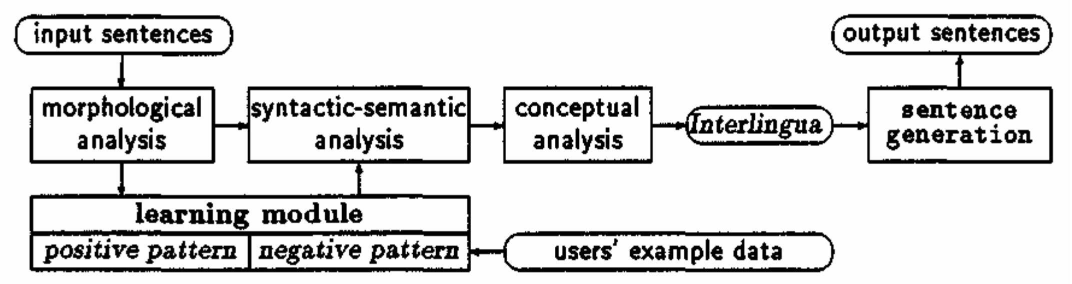


Figure 7: A MT system configuration with the learning module

The learning module memorizes two different kinds of patterns: adequate relationships as positive patterns and inadequate relationships as negative patterns. The module is capable of selecting an adequate structure as well as disambiguating inadequate structures.

3. Pattern representation by the salient features

   The patterns are represented in the form of five regions by the salient features. The model is able to memorize the global structures composed of any number of clauses, because the other regions than a modifier clause region and a modifiee clause region are represented as three context-determinants: a pre-clause region, an inter-clause region and a post-clause region. All regions are represented by the salient features. The features are proved to control the global structures by indicating clause relationship preference. Indeed, they are used for describing heuristic rules in the MT system. They are valid for representing the global structure. The features can be automatically induced from the morphological analysis result of the input sentences, based on the lexical information in the dictionary. They are practically extractable.

5 Preliminary Experiments by a Connectionist

Some connectionist approaches are applied for structure learning like context understanding and syntactic-semantic analysis [11, 23, 24, 25, 26, 27]. There are the following properties.

1. Context understanding
   Some events are linked with each other by stimulant and repressive relation. A connectionist learns a kind of causality and adjacency between two events, and offers a converged result. Some rules are implicitly learned according to the causality and adjacency by the connectionist, and such implicit rules are necessary for determining the structure. Therefore, a connectionist approach is more prospective than other corpus-based approaches.

2. Syntactic-semantic analysis
   It is proposed to learn syntactic and semantic relationship between a predicate and its case element according to their semantic features. Between them, there is not any causality and adjacency, but some constraints. Therefore, it is required to prepare a lot of examples for obtaining more information than described in the dictionary as constraints [11].

Analysis of a global structure among clauses has similar property to the context understanding. The clause corresponds to an event of the context understanding. Between two clauses, there is a logical preference like causality and adjacency of the context understanding. Therefore, a connectionist approach can be expected for the global structure learning. The approach is also preferable because it enables a human to arrange the connectionist network and to control weights of the linkage according to linguistic expectation.

The authors made some preliminary experiments to confirm the possibility of learning the global structures of compound sentences consisting of five clauses, by using simple feed forward neural network with a three-layer perceptron. The experiments revealed that 1256 relationships between two clauses were learned from the sentences, and that 94% of them could be recollected from the network[28]. They suggested that the model was effective for global structure selection according to users' examples.

6 Conclusion

The authors proposed a hybrid model for analyzing the global structures by taking advantage of users' examples. This model has three characteristics: a hybrid analysis model, a positive and negative pattern learning, and pattern representation by the salient features. They enable the model to be integrated in the MT systems by solving the problems for memorizing structures. The model was investigated in comparison with the properties of connectionist approaches, and then the validity of the model was supported by the the results of preliminary experiments.

After the model is verified and improved by a larger and more various set of compound sentence examples, it will be implemented for the practical use in the MT system.

Acknowledgment

The authors thank Mr. Ken-ichi ISO for his suggestions regarding a learning mechanism, and Mr. Shin-ichiro KAMEI for his valuable comments on a *Lexical Discourse Grammar.* They also thank Mr.Masao WATARI for his advice and continuous supports.

References

[1] Kamei,S. and Muraki,K. "A Proposal of Lexical Discourse Grammar," *Proc. WGNLC of the IECIE,* Vol.86 No.189 NCL86-7 ,1986 (in Japanese).

[2] Doi,S., Muraki,K. and Kamei,S. "Lexical Discourse Grammar and its Application for Decision of Long Distance Dependency (II)," *Proc. WGNLC of the IECIE,* NCL91-29(PRU91-64),1991 (in Japanese).

[3] Fujii,Y., Suzuki,K., Maruyama,F. and Dasai,T. "Analysis of Long Sentence in Japanese-English Machine Translation System," *Proc. Information Processing Society of Japan,* 1F-1, May 1990 (in Japanese).

[4] Brown,P., Cocke,J., Della Pietra,V, Jelinek,F., Mercer,R. .and Roossin,P. "A statistical approach to language translation," *Proc. of Coling* '88, pp.71-76, 1988

[5] Rimon,M.,McCord,M.,Schwall,U.,and Martinez,P. "Advances in Machine Translation Research in IBM," *Proc. of Machine Translation Summit III*,pp.11-18,July 1991

[6] Nagao.M. "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle," *Artificial and Human Intelligence*, Elithorn & Banerji, Eds., pp. 173-180, North-Holland, Elsevier Science Publishers, 1984

[7] Victor Sadler, "An example of (simulated) machine translation based on a Bilingual Knowledge Bank," *MT based on Analogy,ATR, Kyoto,Japan*, BSO,2BI-2B17, June 1990

[8] Sato,S. and Nagao,M. "Toward Memory-based Translation," *Coling-90,* pp.247-252. 1990

[9] Sumita,E., Iida,H. and Kohyama,H. "Translating with Examples: A New Approach to Machine Translation," *The Third Inte'l Conf. on Theoretical and Methodological Issues in Machine Translation of Natural Language*, 1990

[10] Sumita,E., Iida,H. and Kohyama,H. "Example-based Approach in Machine Translation," *InfoJapan'90,* Part 2 pp.65-72, 1990

[11] Murase and Nakagawa "A kakari-uke analysis of bunsetu lattice using Boltzmann-machine," *Proc. Information Processing Society of Japan,* 5E-8, pp.947-948,May 1989 (in Japanese).

[12] Nishida.T. "Connectionist Model and its Application to Natural Language Processing," *Proc. Symposium on Natural Language Processing,* IPSJ, January 1988 (in Japanese).

[13] Wilks,Y., Huang,X. and Fass,D. "Syntax, Preference and Right Attachment," *Proceeding of IJCAI-85*, pp779-784,1985

[14] Sergei Nirenburg "Trends in Knowledge-Based Machine Translation," *International Symposium on Multilingual Machine Translation '90*, pp.41-45, Tokyo,Japan, November 1990

[15] Carbonell,J.G. and Tomita,M. "Knowledge-Based Machine Translation,The CMU Approach. In S.Nirenburg (ed.)," *Machine Translation: Theoretical and Methodological Issues* .Cambridge University Press,pp.68-89

[16] Carbonell,J.G., R.E.Cullingford, and A.V.Gershman "Steps toward Knowledge-Based Machine Translation," *IEEE Translation on Pattern Analysis and Machine Intelligence*,July. 1981

[17] Nagao,K. "Dependency Analyzer: A Knowledge-Based Approach to Structural Disambiguation." *Coling-90*, pp.282-287. 1990

[18] Russel,G.,A.Ballim,D.Estival,and S.Warwick "A Language for the Statement of Binary Relations over Feature Structure," *Proc. of the 5th Conference of the European Chapter of ACL*, pp.287-292, April 1991

[19] K.Muraki "VENUS: Two-phase Machine Translation System," *Future Generations Computer Systems*, 2, pp. 117-119, 1986

[20] Okumura,A., Muraki,K. and Akamine,S. "Multi-lingual Sentence Generation from the PIVOT interlingua," *Proc. MT SUMMIT III*, pp.67-71, July 1991

[21] Okumura,A. and Muraki,K. "French Sentence Generation from the PIVOT interlingua," *Proc. Pacific Rim International Conference on Artificial Intelligence*, September 1992 (to be appeared)

[22] Stanfill,C. and Waltz,D. "Toward Memory-Based Reasoning," *Comm. of ACM*, Vol.29, No.12. pp.1213-1228,1986

[23] Selman Bart and Hirst Graeme "A rule-based connectionist parsing system," *Proceedings of the Seventh Annual Conference of the Cognitive Science Society, Irvine, CA,* pp.212-221, August 1985

[24] Waltz,D.L. and Pollack,J.B. "Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation," *Cognitive Science,* Vol.9 pp.51-74, 1985

[25] Charniak, E. "A Neat Theory of Marker Passing," *Proc. AAAI-86,* pp.584-588, 1986

[26] McClelland,J.L. and Kawamoto,A.H. "Mechanism of Sentence Processing," *Assigning Roles to Constituents of Sentences,* Vol.2, pp.272-325, 1986

[27] Takahashi,N. and Itahashi,S. "Japanese Modification Analysis with Mutually Linked Neural Network," *Proc. Information Processing Society of Japan,* 4F-7, pp.464-465 May 1990 (in Japanese).

[28] Okumura,A., Yamabana,K. and Muraki,K. "Learning of Japanese Syntactic Dependency by Neural Network," *Proc. of the 3rd Annual Conference of JSAI*, Vol 1, 11-5, pp.353-356, 1990 (in Japanese).