ICOT Technical Report: TR-190

TR-190

RESEARCH ACTIVITIES ON NATURAL LANGUAGA PROCESSING OF THE FGCS PROJECT

by
T. Yokoi, K. Mukai, H. Miyoshi
and Y. Tanaka

June, 1986

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03) 456-3191~5
Telex ICOT J32964

Institute for New Generation Computer Technology

RESEARCH ACTIVITIES ON NATURAL LANGUAGE PROCESSING OF THE FGCS PROJECT

Toshio Yokoi, Kuniaki Mukai, Hideo Miyoshi, Yuichi Tanaka

Institute for New Generation Computer Technology(ICOT)
Mitakokusai Building 21F,
1-4-28, Mita, Minato-ku, Tokyo 108, Japan

ABSTRACT

The research activities on natural language processing of the FGCS Project are presented. Linguistic phenomena are formalized in terms of complex structures and constraints on them. The logic programming paradigm is adopted for implementing natural language processing systems because the basic operation for the complex structures is isomorphic with respect to unification. DUALS (Discourse Understanding Aimed at Logic-based System), CIL (Complex Indeterminate Language), and JPSG (Japanese Phrase Structure Grammar) are being developed using the unification-based approach. The large-scale machine readable and understandable dictionaries are also being developed.

1. Introduction

In our daily life, we communicate with one another mainly by means of speech and writing in natural languages. We get a lot of information from books and papers. Communication between human and computer should also be performed in the medium of natural language. The ability of computers to understand natural language will increase their accessibility and flexibility. The Japanese fifth generation computer project aims to develop such intelligent computer systems.

The special characteristic of the Japanese language, i.e., the use of a great many Chinese characters, has made Japanese text input and processing difficult for a long time. Recently, however, Japanese language processing technology has advanced a lot as evidenced by Japanese word processors and commercial machine translation systems. These technologies combined with artificial intelligence are expected to provide new Japanese information processing technology and a new computer culture.

ICOT began research and development of the Fifth Generation Computer Systems (FGCS) in 1982. Natural language processing technology is one of the most important research themes for the FGCS Project because it is a fundamental technology for knowledge information processing and it is used for the research and development of knowledge-base, intelligent-interface and various basic application systems, such as machine translation systems.

The results of research in the initial stage have led us to the conclusion that the logic programming framework is the most suitable for implementing natural language processing systems [6]. The linguistic phenomena in natural language can be formalized in terms of the complex structures of the grammatical features and the constraints on them, as is seen in the feature set of Generalized Phrase Structure Grammar (GPSG) [7], the functional structure of Lexical Functional Grammar (LFG) [11] and the "dags" of PATR-II [20]. The basic mechanism of logic programming is unification in Horn clause logic. Definite Clause Grammar (DCG) [10] is one of the bridges connecting the natural language processing and logic programming. Most of our research activities can be regarded as the improvement and the extension of DCG. GALOP (BUP) [12] is a bottom-up left corner parser which overcomes the drawbacks of top-down parser for DCG. A parallel model of DCG is also being developed [13]. CIL (Complex Indeterminate Language) [15,16] was developed to express and operate the complex features. CIL is an extension of

/

Prolog. The newly introduced "partially specified term" of CIL is suitable for representing the complex features because an extended unification is defined on two partially specified terms. The declarative constraints can be written using the freeze mechanism of CIL. Our approach for semantic analysis, which also deals with pragmatics, is based on situation semantics [1] theory. In this approach, the semantic analysis process corresponds to constructing the relations between situations, and it is implemented as an algebra of events. The merging of events is one of the basic operations and it has an isomorphic structure with respect to unification. Therefore, the basic operations in both syntactic and semantic analyses are isomorphic with respect to those of logic programming, which makes logic programming compatible with natural language processing. DUALS (Discourse Understanding Aimed at Logic-based Systems) is its application system for discourse understanding which reads stories and answers questions on the stories. JPSG (Japanese Phrase Structure Grammar) [9] is a GPSG-based grammar theory for Japanese language whose basic operation is unification. The unification-based parser for JPSG has been developed. We developed some application systems, in order to verify and evaluate this fundamental technology mentioned above. Finally, the processing of large-scale language data is another important aspect of natural language processing. Dictionaries include much information about syntax and semantics which will be utilized for designing the lexicon and the knowledge-base in natural language processing systems. The following three types of machine readable and understandable dictionaries will be developed in the subproject which started in April 1986:

(1) Basic Word Dictionaries:
Four machine readable master dictionaries with 200,000 entries in each dictionary.
(2) Concept Classification Dictionary:
A systematic dictionary for 400,000 concepts including a general thesaurus.
(3) Concept Description Dictionary:
A knowledge database containing semantic descriptions of 400,000 concepts.

These systems mentioned above are implemented on the personal sequential inference machine PSI [17]. The programs are written in its programming language ESP [4]. This paper describes the main research activities and plans for natural language processing of the FGCS Project -- CIL, DUALS, JPSG, and Machine Readable and Understandable Dictionaries.

2. CIL (Complex Indeterminate Language)

2.1 Partially Specified Term

CIL is an extension of Prolog which was designed for the system description language of DUALS. CIL has the freeze predicate, which was originally introduced in Prolog-II [5], as a primitive predicate for realizing various lazy evaluation controls.

CIL introduces a new type of object called "partially specified term" ("partial term" for brief), which is influenced mainly by the notion of assignment developed in the situation theory of [2,3].

CIL = Prolog + Partial Term + Freeze.

We understand partial term as an abstraction from the following data structures, which are widely seen in programming languages, grammar formalisms, etc.:

-- Herbrand term in first order logic.
-- Association list and property list in LISP.
-- Frame and unit in knowledge representation.
-- Record in programming languages.
-- Record in relational data base theory.
-- Assignment in Barwise's situation theory.
-- Category as complex feature in GPSG and functional structure of LFG.

A partial term is written in CIL like this:

{a1/b1, ..., an/bn} n>= 0

where each ai is a ground term and bi is any term, possibly a partial term. The ordinal unification is extended to the partial terms. For example the extended unifier works like this:

unify({a/1, b/2}, {b/X, c/3})
= {a/1, b/2, c/3},

unifying X to be X = 2.

CIL can represent a semantic network even including cycles by using partial terms. For instance, the CIL unifier solves the system of three equations A = B, A = {a/B}, and B = {a/A}, giving A = B = {a/A}, a singleton graph with a self-loop with an edge labelled a. As is easily seen, CIL unification is close to that over infinite trees in Prolog-II. The domain of CIL can be defined formally to be a set of infinite trees.

2.2 Reserved Forms in CIL

The current CIL syntax is an extension of the syntax of DEC-10 Prolog. The following symbols ':', '?', '@', '#', '??', '!' appearing in terms are reserved for the CIL system as follows:

(1) A term of the form X!a is equivalent as a term to the value of the slot of X whose name is a. That is,

{a1/b1, ..., ai/bi, ..., an/bn}!ai = bi.

(2) A term of the form X:C with terms X and C is called a description. C should be an executable form. This term is read "X such that C".

(3) A literal of the form p(...X?...) is equivalent to the literal freeze(X,p(...X...)).

(4) CIL includes convenient forms of term which are defined as follows:

- ap <=> X:p(X?), where p is a predicate symbol of arity 1 and X is a new variable.
- Vap <=> V:p(V?), where p is like the above.
- V=T <=> V:(V=T).
- V?? <=> X:(X?=V), where X is a new variable.

2.3 Situation Semantics in CIL

Although the current CIL is not a full implementation of situation theory yet, it is already useful because of the introduction of partial terms and extended unification over them. Partially specified terms have general and natural descriptive power to represent various data structures of objects necessary for situation theory. The most difficult and basic problem which remains open for CIL, however, is to develop some ideas for designing a control library for constraint description. We think that the problem corresponds directly to the implementation of the constraints of situation theory.

3. DUALS (Discourse Understanding Aimed at Logic-based Systems)

DUALS is an experimental discourse understanding system developed to build a computational model for discourse understanding. The semantic framework is situation semantics in which the sentence meanings are represented as relations between situations. DUALS aims at dealing with the following items within this framework.

1) Primitives for discourse understanding
   - (a) Anaphora
   - (b) Speech act
   - (c) Attitude verb
   - (d) Tense
   - (e) Quantifier
   - (f) Conditions
   - (g) Coordination
2) Plan goal
3) Type description
4) Predicates for manipulating situations

The latest version of DUALS was implemented in CIL. It reads a story written in Japanese language and answers various type of questions about it. The system has the following characteristics:

(1) The semantic structure is constructed with the objects used in situation theory, such as individuals, assignments, relations, locations, conditions, events, parameters, and so on.

(2) Syntax analysis is performed by the parser based on the concurrent process model called SAX (Sequential Analyzer of syntaX and semantics) [18]. There are about 250 grammar rules. Constraints between

situations representing sentence meaning are generated in the form of partially specified terms of CIL by the syntax analysis module.

(3) Anaphora processing algorithm, i.e. the identification of pronouns and zero-pronouns (ellipses) is based on Kameyama's model [10].

(4) Plan-goal-based discourse structures are obtained by the discourse processing module. The rules to construct the discourse structures are described as constraints between events.

(5) The sentence generation module generates the surface sentences from internal meaning structures using grammar rules.

Our technical approach to implementation is to build a package for extended unification in logic programming. An interesting problem, and a more theoretical challenge, is determining what kinds of unification are needed as primitives for implementing situation semantics.

4. JPSG (Japanese Phrase Structure Grammar)

Grammar is an important component of a system for natural language understanding. JPSG is a new Japanese grammar theory for Japanese language based on GPSG. GPSG is suitable for implementation in the logic programming paradigm because it is a natural language syntax theory based on context free grammar (CFG) and its basic computational mechanism is unification. Besides, GPSG has the following features:

(a) Syntactic categories are defined as a complex feature set.

(b) Only phrase structure is used to represent grammatical information.

(c) Metarules for phrase structure rules are introduced.

(d) Constraints on features are described in the syntactic principles, which make phrase structure rules general.

(e) Syntax and semantics are closely related.

Since the Japanese language has a word order variation called "scrambling", GPSG cannot handle it feasibly. In order to handle the "scrambling", the subcategorization feature (SUBCAT) whose value is a set of syntactic categories, is introduced in JPSG. This is an extension of HPSG [19].

Currently, the grammar formalism of JPSG is completed for basic Japanese syntax with the following characteristics:

(1) Syntactic categories are defined as a feature set. Some feature examples are as follows:

PAS -- This indicates the passivizability of a verb. The value is '+' or '-'.

POS -- This indicates a part of speech. The value is one of {V, N, P,...}.

GR -- This indicates a grammatical relation. The value is one of {SBJ, OBJ}.

SUBCAT -- This is the set of syntactic categories which a head category demands as its complements.

SLASH -- This is the set of the missing categories. This feature is used in the same way as in GPSG.

(2) The following phrase structure rule is sufficient for basic Japanese syntax:

(2.1) M --> D H

Rule (2.1) states that mother category (M) dominates one daughter category (D) on the left and one head category (H) on the right. This simplification of the rule is achieved by describing the constraints on the features in syntactic principles.

(3) Since a new SUBCAT feature is introduced, SUBCAT feature principle (SFP) is extended like that of HPSG. SFP describes the inheritance of SUBCAT values as follows:

"The SUBCAT of M is identical to that of H minus D."

For example, if the category of D is N[OBJ] and SUBCAT of H is {N[SBJ], N[OBJ]}, then SUBCAT of M will be {N[SBJ]}. N[OBJ] represents the object noun phrase. The absence of order between the elements of SUBCAT makes it easy to deal with "scrambling" [8].

(4) Most of the syntactic principles used in GPSG are also used in JPSG such as HEAD feature

convention (HFC) and FOOT feature principle (FFP). A set of certain features is called "HEAD features", for example, the POS feature. HFC is a constraint stating that "HEAD features of M are identical to those of H in (2.1)". SUBCAT and SLASH are called FOOT features. FFP is a principle about the inheritance of FOOT features stating that "FOOT feature of M is identical to the union of that of D and H in (2.1)".

(5) A lot of grammatical information is contained in a dictionary.

The basic operation used in JPSG is "unification" because in the syntactic principles mentioned above, the phrase "be identical to" can be replaced by "can be unified to". The parser for JPSG is being developed in CIL. JPSG and CIL are compatible because the syntactic category as feature set corresponds to a partially specified term and syntactic principles correspond to Horn clauses.

5. Machine Readable and Understandable Dictionaries

We presented the unification-based approach for natural language processing and its applications in previous sections. On the other hand, processing of large-scale language data is another important aspect of natural language processing.

This research aims at developing a large-scale database for various natural language processing and speech processing application systems. The language database will be composed primarily of three machine-readable dictionaries: a large-scale basic dictionary as the master dictionary; a concept classification dictionary including a thesaurus; and a concept description dictionary containing descriptions of the meanings of concepts. Application systems utilizing these dictionaries will be developed including machine translation systems and speech recognition systems.

5.1 Basic Word Dictionaries

The term "basic word" means words used in everyday speech, general technical terms, proper nouns, and so on. Machine-readable master dictionaries will be developed containing these basic words. These are the dictionary types:

(1) Japanese
(2) English
(3) Japanese-English
(4) English-Japanese

Each dictionary will include about 200,000 entry words. These dictionaries will be developed in accordance with the specifications already established [14].

5.2 Concept Classification Dictionary

This dictionary will contain specifications of the relations between concepts and indicate exactly how specific concepts are classified in the concept world. Classification bases for the concept world are 'super-sub', 'whole-part', 'composition-element' and other similar relations. The multiple inheritance mechanism will be used as well. The standard thesaurus will form a part of this dictionary. At least 400,000 concepts will be included.

5.3 Concept Description Dictionary

This dictionary will contain the meaning of each individual concept classified in the concept classification dictionary. The combination of the concept classification and the concept description will form the knowledge base for the "general world", and will be utilized in semantic and discourse analysis.

5.4 Application systems

Machine translation systems and speech recognition systems will be developed using these dictionaries.

6. Conclusion

This paper described the research activities on natural language processing within the Japanese Fifth Generation Computer System project. Having finished the initial stage, the project is now at the end of the first year

in the four-year intermediate stage. Natural language understanding includes a lot of difficult issues that remain unsolved, especially in discourse understanding. Nevertheless, fruitful results and new ideas have been obtained over the four years of research to date by concentrating on the logic programming framework as discribed in this paper. The last four years has convinced us that the logic programming approach is very promising for implementing natural language processing systems. In the intermediate stage, we will continue this approach to build the subsystems that will be integrated to form the total knowledge information processing system in the final stage.

7. Acknowledgement

This research is being carried out by the Second Research Laboratory of ICOT in very close cooperation with seven manufactures. Many fruitful discussions were held in meetings of the following working groups -- NLS (Natural Language processing System, Head: Prof. Tanaka of Tokyo Institute of Technology), JPS (Japanese Phrase Structure grammar, Head: Prof. Gunji of Osaka Univ.), MRD (Machine Readable Dictionaries, Head: Prof. Ishiwata of Ibaraki Univ.). Finally, the authors would like to thank the director K. Fuchi, ICOT Research Center for providing us with the opportunity for this research.

[References]

[1] Barwise, J. and Perry, J., Situations and Attitudes, MIT Press, 1983.

[2] Barwise, J., The Situation in Logic-III: Situations, Sets and the Axiom of Foundation Center for the Study of Language and Information, CSLI-85-26, 1985.

[3] Barwise, J., Recent Developments in Situation Semantics, Proc. of International Symposium on Language and Artificial Intelligence, 1986.

[4] Chikayama, T., ESP Reference Manual, ICOT TR-044, 1984.

[5] Colmerauer, A., Prolog-II: Reference Manual and Theoretical Model, Internal Report, Groupe Intelligence Artificielle, Universite d'Aix-Marseille II, 1982.

[6] Furukawa, K. and Yokoi, T., Basic Software System, Proc. of FGCS'84, 1984.

[7] Gazdar, G., Klein E., Pullum G., and Sag I., Generalized Phrase Structure Grammar, Oxford, Basil Blackwell, 1985.

[8] Gunji, T., Subcategorization and Word Order, Proc. of International Symposium on Language and Artificial Intelligence, 1986.

[9] Gunji, T., Japanese Phrase Structure Grammar, D. Reidel Publishing Company, (to appear).

[10] Kameyama, M., Zero Anaphora: The Case of Japanese, Draft of Ph.D Diss., Dept. of Linguistics, Stanford Univ., 1984.

[11] Kaplan, R. and Bresnan, J., Lexical Functional Grammar: A Formal System for Grammatical Representation, in Mental Representation of Grammatical Relations (Bresnan eds.), MIT Press, 1982.

[12] Matsumoto, Y., Tanaka, H., Hirakawa, H., Miyoshi, H. and Yasukawa, H., BUP: A Bottom-Up Parser Embedded in Prolog, New Generation Computing, OHMSHA, LTD. & Springer-Verlag, Vol.1, No.2, 1983.

[13] Matsumoto, Y., A Parallel Parsing System for Natural Language Analysis, Proc. of the 3rd ICLP, 1986.

[14] Miyoshi, H. Tanaka, Y., Yokoi, T., Ishiwata, T., Tanaka, H., Amano, S., Uchida, H. and Ogino, T., Basic Specifications of the Machine-Readable Dictionary, ICOT TR-100, 1985.

[15] Mukai, K., Horn Clause Logic with Parameterized Types for Situation Semantics Programming, ICOT TR-101, 1985.

[16] Mukai, K., Unification over Complex Indeterminates in Prolog, ICOT TR-113, 1985.

[17] Nishikawa, H., Yokota, M., Yamamoto, A., Taki, K., and Uchida, S., The Personal Inference Machine (PSI): Its Design Philosophy and Machine Architecture, ICOT TR-013, 1983.

[18] Pereira, F. and Warren, D.: Definite Clause Grammar for Language Analysis - A Survey of the Formalism and a Comparison with Augmented Transition Networks, Artificial Intelligence, 13, 231-278, 1980.

[19] Pollard, C., Lecture on HPSG, Unpublished Lecture Notes, Stanford University, 1985.

[20] Shieber, S. M., Using Restriction to Extended Parsing Algorithms for

Complex-Features-Based Formalisms, Proc. of
23rd ACL, 1985.