PROCEEDINGS OF THE INTERNATIONAL CONFERENCE
ON FIFTH GENERATION COMPUTER SYSTEMS 1988,
edited by ICOT. © ICOT, 1988

Overview of the Dictionary and Lexical Knowledge Base Research

Yuichi Tanaka Tsutomu Yoshioka
Institute for New Generation Computer Technology
Mita-Kokusai Bldg 21F, 4-28, Mita 1-chome, Minato-ku
Tokyo 108 JAPAN

ABSTRACT

Overview of structure and contents of our dictionary and lexical knowledge base is described in this paper.

There are two main objectives in ICOT's research on natural language processing, *i.e.* experimental research on discourse understanding and development of efficient, general purpose natural language processing system.

Discourse understanding systems based on logic programming have been investigated since 1982 at ICOT. The focus of the research has been mainly on basic mechanisms of discourse understanding, and the parallel algorithms for these mechanisms.

Experimental system called *DUALS* has been developed and to verify the research results.

These results have been put together to form a general purpose software environment for natural language processing named *LTB* (Language Tool Box).

This paper presents an overview of the dictionary and lexical knowledge base for both *DUALS* and *LTB* , and discusses the future research and development direction of lexical knowledge bases.

1 Introduction

It might be said that Natural Language Processing (NLP) has come to a turning point. In the first place, we have come to the situation in which we have to develop new ideas and mechanisms to deal with large amount of data because current NLP, such as machine translation, is involved in the stream of vast data in the real world. Secondly, to qualify the NLP, we need much deeper semantic analysis such as sentence/discourse understanding, inference of speaker's/hearer's mental state, and so on. In other words, we must increase the quantity and quality of NLP to get over today's difficult situation. It has been recognized that large-scale dictionary or lexical database with precise semantic description is indispensable to support NLP of today and further.

In ICOT, research and development of general purpose Japanese language processor (LTB) has made the way of high speed processing of vast amount of actual data, as well as research experience of discourse understanding experiment system (DUALS) has shown clues to establishment of mathematical basis of discourse understanding.

We are developing a dictionary and lexical database for these two purposes. We have developed a dictionary with several thousand entries of Japanese words for the LTB system, and are planning to increase the vocabulary to ten times larger in the near future. At the same time, we have prepared a dictionary for DUALS which includes 2,000 entries with precise semantic description. One of our objectives is to integrate these two dictionaries to produce a semantic dictionary of appropriate size.

2 Common Tool vs World Knowledge

In the very early stage of NLP, researchers developed experimental small dictionaries for their own use. Next, printed dictionary for human use was substituted as a source to reproduce NLP databases, and now large-scale dictionaries for strict computer purposes are developed from scratch again by many institutions in the world. Main reason is that deep semantic analysis in current NLP requires various new information which cannot be extracted from traditional dictionaries, in which complicated description of meaning is, for instance, replaced by example sentence in many cases.

In NLP, highly sophisticated inference based on so called "common sense" or world knowledge is necessary to understand the meaning of sentence/discourse, to analyze sentences with infering speaker's intention, or to generate sentences with guessing hearer's mental state. In order to support this kind of inference, world knowledge must be described in the dictionary.

World knowledge base of this kind can be used as a basic database not only in NLP but also in the variety of knowledge information processing systems. Actually, the trial has been done in which the articles of an encyclopedia are transformed into some terms in a knowledge representation language. Its apparent target is to get over vulnerability of "expert" knowledge information processing without world knowledge. However, in a long period, it aims to acquire further knowledge by learning, and to produce an intelligence by combining its own knowledge.

World knowledge being also indispensable for our research, we must accumulate sufficient knowledge into our dictionary. Distinction between lexical knowledge and world knowledge is a problem here. How to express both knowledge is another problem. We will propose a framework to deal with these problems.

It is difficult to distinguish world knowledge from lexical knowledge because the boundary is vague. We have decided to select following items as lexical knowledge: morphological information, syntactical information, case information, semantic feature and cooccurrence relation. Other items are involved in world knowledge. They are expressed in a constraint style description, which will be shown precisely in the later section.

3 Overview of the Dictionary

We are developing the lexical knowledge base for the both objectives described above. It is necessary to construct two different dictionaries together, each of which is explained in the following subsections.

3.1 The LTB dictionary with lexical knowledge

We have already completed this type of dictionary with 4,000 entries. This is used for *LTB* (general purpose japanese language processor) system, where each subsystem (lexical analyzer, syntax analyzer, sentence generator) transforms this dictionary into its own internal representation. The dictionary includes lexical knowledge such as morphological information, syntactical information, case information and basic semantic information.

The semantic information is expressed in basic soa (state of affairs) form, called *infon* in terms of situation theory and situation semantics. The syntax analyzer *SAX* makes use of this information in the way that it composes the semantics of a sentence out of each infons corresponding to each surface morphemes according to the dependency structure of a sentence, at the same time it solves cooccurrence constraints of the surface words in order to reduce ambiguity.

In our grammar, we established 7 parts of speech listed below:

noun (*meishi*)
: Stems for both *sahen-meishi* and *keiyou-doushi* are also included in this category. Prefixes and postfixes are considered to be nouns with complement.

verb (*doushi*)
: Words which conjugate and have *u*-ending in present form.

adjective (*keiyoushi*)
: Words which conjugate and have *i*-ending in present form.

adverb (*fukushi*)
: Words which do not conjugate, do not appear at the end of sentences, and are able to modify verbs and adjectives.

nominal modifier (*rentaishi*)
: Words which do not conjugate and modify nouns only.

auxiliary verb (*jodoushi*)
: (Our dictionary does not include auxiliary verbs, which are treated by grammar rules instead.)

particle (*joshi*)
: (Our dictionary does not include particles, which are treated by grammar rules instead.)

Dictionary format, which will be described in section 4, is designed appropriately for each part of speech.

Currently we are expanding vocabulary up to 100,000 entries. This supplement part contains only morphological information and some items in syntactical information which can be used for *LAX* system.

3.2 DUALS dictionary with world knowledge

We have given the detailed semantic information to 2,000 words within *LTB* dictionary. These words appear in the evaluation text of *DUALS* version 3.

As explained above, world knowledge is described in constraint form. Word meaning consists of two parts, the first of which will be a direct constituent of a sentence meaning called *infon*, the second of which controls validity and applicability of infon called *constraint*.

Specific constraints are described in each entry of semantic dictionary, on the other hand, general constraints are stored in the constraint dictionary. Constraint solver deals with these constraints impartially.

4 Structure and Contents

The dictionary consists of three parts, entry word dictionary, semantic dictionary, and constraint dictionary. In this section, structure and contents of entry word dictionary and semantic dictionary are described.

4.1 Entry Word Dictionary

The structure of entry word dictionary is common to all part of speeches. The entry word dictionary consists of entry word records, shown in Fig. 1, prepared for each word.

Currently over 4,000 records are used by *LAX* and *DUALS* system.

4.2 Semantic Dictionary

The semantic dictionary consists of semantic records, the structure of which is peculiar to each part of speeches. The structure of semantic record of verbs is shown in Fig. 2.

Average number of meaning within an entry being over 4, the number of semantic records is nearly 20,000. These records are compiled together with entry word records, and used by *LAX* and *DUALS* system.

ID : Unique identification number for each word.

Part of speech : Part of speech listed in the previous section.

Conjugation type : For conjugated words (*i.e.* verbs and adjectives), conjugation types are specified. For other words, this slot is omitted.

Surface expression : Surface expression is described in *kanji* and *kana* characters. For conjugated words, expression is divided into stem and ending.

Pronunciation : Pronunciation is described in *kana* characters. Variant pronunciations are also listed.

Semantic index : List of link pointers to the semantic dictionary. For words having more than one meaning, each element of the list points each different meaning. Difference of deep case combination is regarded as defference of meaning.

Figure 1: Format of Entry Word Record

5 Semantic Description

5.1 Classification of Words

In describing semantics of words, it is necessary to classify words into their function class first. We have assumed the following three classes:

ID : Unique identification number for each meaning. The semantic pointer in the entry word record above uses this ID number.

Deep case : Deep case indices of a verb. Optional cases are ignored.

Active surface case : In case of active voice, pairs of deep case index and surface case particle are listed.

Passive surface case : In case of passive voice, list of pairs of deep case index and surface case particle are described. If there are more than one passive voice expressions to active one, alternative lists are also described. If there are no passive voice expression, this slot is omitted.

Arguments : Correspondence between deep case index and argument in the semantic structure is described.

Semantic structure : Semantic structure of verb in an infon style.

Constraints : Constraints for arguments in logical expression.

Thesaurus codes : Thesaurus codes for verb.

Complement : In case of function verbs, which must take verb or noun complement to form the meaning, the complement class are described here.

Voice : Existence of four expressions related with voice, *i.e.* direct passive, indirect passive, causative, giving-receiving expression.

Aspect : Original aspect of the verb alone.

Transitivity : Distinction between transitive and intransitive verb.

Volitivity : Volitional or not.

Figure 2: Format of Semantic Record

Function Words: Auxiliary words, function nouns, function verbs. These words attach to other words to produce new meaning or to transform the original meaning. They have no concrete meaning but transformational function. The semantics of these words can be described in syntactical level.

Basic Words: Adjectives, adverbs, basic nouns, basic verbs. These words have basic and abstract meaning, and are used very widely. Metaphorical usage is available. There being only less than 1000 adjectives in Japanese, the semantics of adjectives are usually abstract and vague. For example, *takai* means high, tall, expensive, etc. Our current objective is to describe the semantics of basic nouns and basic verbs.

Other Words: Words which have concrete and compound meaning. Most nouns and Cino-Japanese verbs are included in this category. However, their syntactic function in sentences is simple. The core of the description of these words is knowledge representation.

Assuming two categories *function nouns* and *function verbs*, we selected words in these categories. Number of function words is several hundred.

Group	Subgroup	Examples
時間	前	前, 以前, 先(さき), …
(Time)	後	後, 以後, 以降, …
	間	間, 合間, 内, …
	最中	最中, 盛り, さなか, …
	始め	始め, 始まり, 当初, 端, …
	終り	終り, 最後, 終局, 終盤, 結び, …
	…	…
場所	上	上, 上側, 上方, 上部, …
(Location)	下	下, 下側, 下方, 下部, …
	前	前, 手前, 先(さき), 表, …
	後	後ろ, 後部, 後方, 背後, …
	間	間, 中, 中間, 区間, …
	…	…
順序	前	前, 先頭, …
(Order)	後	後ろ, 後尾, …
	…	…
程度	上	上, 以上, …
(Degree)	中	中, 半ば, 中程, …
	下	下, 以下, …
	…	…

Figure 3: Example of Function Nouns

There are about 4000 *basic nouns* and *basic verbs* in Japanese. We have classified the meaning of basic words by the thesaurus which was created anew for our purpose. For verbs, we indicate explicitly the relation between surface representation, semantic structure and semantic constraints.

5.2 Function Nouns

Japanese nouns cover very large range of the semantic world. We count as function nouns the nouns which have no correspondence to reality and can express meaning when connecting to other words or phrases. In this definition, we focus on the function of a word in a sentence. For example, the noun *mae* (front) means nothing in the real world, however, it inputs an object to output location if it is used in the form *X no mae* (in front of *X*).

There are so many kind of function nouns, such as *time*, *location*, *order*, *degree*, etc. Examples of function nouns are shown in Fig. 3.

object

- inanimate object
  - artificial object
    - material
      lumber, steel
    - parts
      tile, battery, IC
    - tools [*use*]
      household
      construction
      communication
      medicine
      …
    - tools [*material*]
      steel
      wood
      plastic
      …
    - tools [*feature*]
      electric
      blade
      …
    - *etc.*

Figure 4: Example of Thesaurus

5.3 Thesaurus for Nouns

We have created a thesaurus which expresses *super-sub* relationship among concepts. This thesaurus is used for cooccurrence checking of verbs and nouns, helping to paraphrase texts, selecting appropriate word in sentence generation, and so on. It has a tree-like structure in which each concepts is represented by a node while super-sub relation is represented by an arc. Each node in the thesaurus has a code which represents the location in the tree. The code is assigned for each word in the dictionary.

Generally in thesaurus, super concepts are divided into some lower concepts that should be mutually exclusive. In conventional thesauri, however, lower concepts are not strictly exclusive, but simply collection of words. To make the division clear, we have introduced *axis of division* or *viewpoint*. When we divide a super concept, first we set up an axis of division, along which some values from that viewpoint are taken. For example, taking *material* as an axis, a concept *tools* is divided into *metal-tools*, *wood-tools*, *plastic-tools*, and so on. On the other hand, the same concept is divided into *kitchen-tools*, *carpenters-tools*, etc. when we take *use* as an axis. A concept is, as a result, usually divided in some higher dimensional space in our thesaurus. Our coding reflects the structure of this space.

This thesaurus is used for cooccurrence checking of the surface words, especially verb and noun combination. This is performed with the semantic constraint information of each cases of verbs. Thesaurus code is involved in this semantic constraint.

Finding higher abstract concept, our thesaurus will help to paraphrase text. This is performed by searching the instances of super concept of a given word. Rhetoric expressions could also be analyzed by similar procedure.

And besides, to reduce the searching space of words to be selected in sentence generation, the thesaurus is useful. In other words, it is difficult to find an appropriate word directly from the formal semantic description, however, one can restrict the candidates with going down the tree structure of this thesaurus. In this case, description with viewpoint is also useful.

A part of our thesaurus is shown in Fig. 4.

5.4 Function Verbs

We have defined function verbs that work merely as syntactical elements carrying no substantial meaning. Verbs or nouns combining with function verbs express the meaning, that will be transformed or modified by function verbs. By this definition, there are many function verbs of various levels, where meaninglessness varies from zero to some extent. For example, in sentences

> ... *oto ga suru*
> sound *subj.* do
> (= it sounds ...)

and

> *kekkon-shiki wo ageru*
> wedding *obj.* raise
> (= to hold a wedding ceremony)

the verbs *suru* and *ageru* are function verbs.

Verbs which express syntactic feature such as voice or aspect are also classified in function verbs by our definition. They are listed in Fig. 5 and Fig. 6.

Voice	Verb	Examples
Mutual	*kawasu* (exchange)	*keiyaku* (contract) *wo kawasu* *aisatsu* (greeting) *wo kawasu* *houyou* (embrace) *wo kawasu*
Basic	*okonau* (do)	*undou* (exercise) *wo okonau* *kouen* (lecture) *wo okonau* *kaiten* (rotation) *wo okonau*
Passive	*koumuru* (suffer)	*higai* (damage) *wo koumuru* *meiwaku* (trouble) *wo koumuru*

Figure 5: Example of Function Verbs (1)

Aspect	Verb	Examples
Inchoative	*hajimaru*	*oinori* (prayer) *ga hajimaru*
	hajimeru (begin)	*kenkyuu* (research) *wo hajimeru* *giron* (debate) *wo hajimeru*
Completion	*owaru*	*shokuji* (meal) *ga owaru*
	oeru (finish)	*shirabe* (investigation) *wo oeru* *giron* (debate) *wo oeru*
Continuation	*tamotsu* (keep)	*chinmoku* (silence) *wo tamotsu* *sesshoku* (contact) *wo tamotsu*
	tsuduku	*sentou* (battle) *ga tsuduku*
	tsudukeru (continue)	*kenkyuu* (research) *wo tsudukeru* *susurinaki* (sob) *wo tsudukeru*
Reiteration	*kurikaesu* (repeat)	*hentou* (response) *wo kurikaesu* *jikken* (experiment) *wo kurikaesu*

Figure 6: Example of Function Verbs (2)

5.5 Deep Case System for Verbs

We have prepared a deep case system to describe case relation of predicate. The deep case information is appeared in a dictionary entry as:

1. deep case
   List of deep cases of the verb. Optional cases are omitted.

2. surface case of active voice
   List of pairs of case index and particle of active voice. Correspondence between deep case index and surface case particle is expressed.

3. surface case of passive voice
   List of pairs of case index and particle of passive voice.

4. semantic constraint on nouns
   List of pairs of case index and thesaurus code for nouns. Nouns being able to appear in the case are restricted by thesaurus code.

5. argument of semantic structure
   List of pairs of case index and argument of semantic structure. Semantic structure of the word has some variables in the expression. These variables are linked to deep case indices by this list.

The example of deep case information of verb is shown in Fig. 7.

- surface expression: 与え・る (give)
- pronunciation: あたえる (*ataeru*)
- part of speech: verb
- deep case: (AGT, GOA, OBJ)
- surface case: (AGT: *ga*, GOA: *ni*, OBJ: *wo*)
- passive case: (GOA: *ga*, AGT: *ni*, OBJ: *wo*),
  (OBJ: *ga*, AGT: *niyotte*, GOA: *ni*)
- thesaurus code:
  (AGT: 1.1/1.2.2, GOA: 1.1/1.2.2, OBJ: 1.2)
- argument: (AGT: a, GOA: x, OBJ: y)

Figure 7: Example of Deep Case Information

5.6 Semantic Structure for Words

Semantic structure of a word is expressed with *basic state of affairs* or *infon* in terms of situation theory and situation semantics. Basically speaking, predicate represents some relation among objects in the world, while noun represents a object having some relation between other objects. These relationship can be represented by infons.

An object is not always able to appear in an arbitrary argument place of infons. Infons sometimes have some relation to other infons such as *equivalence*, *involve*, etc. This kind of information is expressed by *constraints* connected to each infons. Constraints such as cooccurrence restriction are described within dictionary entry for each word, while general constraints between infons (*e.g.* "Kissing means touching.") are stored in a constraint dictionary.

6 Conclusion

Current subjects on the dictionary and lexical knowledge base research are as follows:

- Extending vocabulary of dictionary. To make our dictionary useful, we must extend the vocabulary up to a hundred thousand at least. We are planning to develop morphological information of this size.

- Describing more knowledge on our lexical knowledge base. The more knowledge it has, the more knowledge it learns.

- Establishing the framework of formal description of tense and aspect. They must be involved in the semantic structure of words, however, current description is insufficient. This research cannot be performed within dictionary system alone. Cooperative study with logical inference module is indispensable.

- Studying rules or mechanism of change of meaning according to change of aspect. Semantic description and deep case information for each word cannot be independent from aspectuality. Then, data in the dictionary should be modified according to the surface word form when applied to the expression in sentences.

To improve our dictionary and lexical knowledge base in both quality and quantity, they should be evaluated with vast amount of actual data. We will make use of *DUALS* and *LTB* as evaluation systems for our dictionary and lexical knowledge base as well as users of them.

ACKNOWLEDGEMENTS

We would like to thank Professor Hozumi Tanaka at Tokyo Institute of Technology and the members of working group NLU/SG-2 for their intensive discussions and suggestions. We would also like to thank Mr. Oshima, Miss Nagasawa, Mr. Umino and Mrs. Kunieda for their enormous effort of dictionary description and theoretical advice, Dr. Shun-ichi Uchida, the chief, and all the members of the Second Research Laboratory for their valuable discussions. We also wish to express our thanks to Dr. Kazuhiro Fuchi, Director of ICOT Research Center, who, as the pioneer in this research area, provided us with the opportunity to pursue this development in the Fifth Generation Computer Systems Project at ICOT.

References

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

[2] Barwise, J. "The Situation in Logic-III, — Situations, Sets and the Axiom of Foundation", CLSI Report No. CLSI-85-26, 1985.

[3] Cohen, P. R., *et al.* "Persistence, Intention and Commitment", CLSI Report No. CLSI-87-88, 1987.

[4] Ikeda, T., *et al.* "Sentence Generation in LTB, *in Japanese*", *5th Conference Proceedings of Japan Software Science and Technology*, 1988.

[5] Kimura, K., Sugimura, R., Takizuka, T. and Mukai, K. "Danwa Rikai Jikken System DUALS dai 2-han no Sekkei to Jissou (Design and Implementation of Discourse Understanding System DUALS-V2, *in Japanese*)", *Proceedings of the 3rd Conference of Japan Society for Software Science and Technology*, Tokyo, 1986.

[6] Matsumoto, Y. and Sugimura, R. "Koubun Kaiseki System SAX no tameno Bunpô Kijutu Gengo (Grammar Description Language for The SAX Parsing System, *in Japanese*), *5th Conference Proceedings of Japan Society for Software Science and Technology*, Tokyo, 1988.

[7] Morioka, K. "Goi no Keisei (Formation of a vocabulary, *in Japanese*)" *Gendai-go Kenkyuu Series 1*, Meiji-Shoin, 1987.

[8] Mukai, K. "A system of Logic Programming for Linguistic Analysis Based on Situation Semantics", *Proceedings of the workshop on semantic issues in human and computer languages.*, CSLI, 1987.

[9] Muraki, S. "Nihongo no Kinou-doushi Hyougen wo Megutte (On Function Verb Expression in Japanese, *in Japanese*)", *Research Report 2*, National Language Institute, 1980.

[10] Ogino, T. "Nihongo no Imi Bunrui Shian (Proposal on Semantic Classification of Japanese Words, *in Japanese*)", *Proceedings of the 31st Conference of Japanese Association for Metrical Linguistics*, 1987.

[11] Okutsu, K. "Seisei Nihon Bunpou-ron (Theory of Generative Japanese Grammar, *in Japanese*)", Taishuukan-Shoten, 1974.

[12] Sano, H., Akasaka, K., Kubo, Y., and Sugimura, R. "Go-Kousei ni Motoduku Keitaiso Kaiseki (Morphological Analysis with derivation and inflection, *in Japanese*), *Proceedings of the 36th Conference of Information Processing Society of Japan*, 1988.

[13] Sugimura, R., Akasaka, K., Kubo, Y., Sano, H., and Matsumoto, Y. "Ronri-gata Keitaiso Kaiseki LAX (Logic Based Lexical Analyzer LAX, *in Japanese*)", *Proceedings of the Logic Programming Conference '88*, ICOT, 1988 (English version will be appeared in *The Lecture Notes on Computer Science.*)

[14] Sugimura, R. "Ronri-Gata Bunpô ni okeru Seiyaku Kaiseki (Constraint Analysis on Logic Grammars, *in Japanese*)", *Proceedings of the 2nd Annual Conference of Japanese society for Artificial Intelligence*, Tokyo, 1988.

[15] Sugimura, R., Hasida, K., Akasaka, K., Hatano, K., Kubo, Y., Okunishi, T., and Takizuka, T. "A Software Environment for Research into Discourse Understanding Systems", *in this proceedings* FGCS'88, ICOT, 1988.

[16] Takizuka, T., Tanaka, Y., and Sugimura, R. "LTB Master Jisho no Kousei (Configuration of LTB Master Dictionary, *in Japanese*)", *Logic and Natural Language Research Group* in Japan Society for Software Science and Technology, 1988.

[17] Miyoshi, H., Tanaka, Y., Yokoi, T., *et al.* "Basic Specification of the Machine-Readable Dictionary", TR-100, ICOT, 1985.

[18] Tanaka, Y., Oshima, M., and Oshima, , R. "LTB Master Jisho no Imi Kijutsu no Kousou (Design of Semantic Description for LTB Master Dictionary, *in Japanese*)", *Logic and Natural Language Research Group* in Japan Society for Software Science and Technology, 1988.

[19] Tanaka, Y., Oshima, M., and Nagasawa, Y. "LTB Master Jisho no Imi Kijutsu (On Semantic Description for LTB Master Dictionary, *in Japanese*)", *Proceedings of the 37th Conference of Information Processing Society of Japan*, 1988.

[20] Tanaka, Y., Umino, B. "LTB Master Jisho no Kouzou to Naiyou (Structure and Contents of LTB Master Dictionary, *in Japanese*)", *Proceedings of the 37th Conference of Information Processing Society of Japan*, 1988.