Unknown Linguistic Knowledge Acquisition System

- Proposal and Implementation of a Prototype System for Japanese Written-Language Document -

Hiroyuki KAMEDA(1) and Chiaki KUBOMURA(2)

(1) Department of Information Technology, Tokyo University of Technology,
Postal address: Katakura 1404-1, Hachioji, Tokyo 192-0982, JAPAN
E-mail: kameda@cc.teu.ac.jp
Tel: +81-426-37-2445 Fax:+81-426-37-2445
(2) Department of Aesthetics & Health Science, Yamano College of Aesthetics,
Poastal address: Yarimizu 530, Hachioji, Tokyo 192-0396, JAPAN
E-mail: kubomurac@yamano.ac.jp
Tel: +81-426-77-2111 Ext.226 Fax: +81-426-77-0234

Summary Advanced natural language (NL) processing systems must have abilities to acquire unknown linguistic knowledge (unknown words and unknown syntactic rules) automatically, so that the systems can improve their linguistic processing performance gradually and developmentally like human children. In this paper, we propose an unknown linguistic knowledge acquisition system based on both a rule-based method and an ILP (Inductive Logic Programming) method. At first, unknown expressions are defined and classified from the NL processing point of view. Principles and a prototype system of unknown linguistic knowledge acquisition are also described in some details. Experiments were also made to evaluate the validity of the system with some amounts of text data resources, which showed fundamental validity of the system.

Keywords: knowledge acquisition, acquisition algorithm, acquisition model

1. Introduction

Techniques of natural language (NL) processing have been making a remarkable progress since the birth of electronic computers, and are now applied to wide ranges of areas of practical engineering. For example, in Japan, word processors are widely spread and used from home to home and from the young to the old, as if they were kinds of stationery, and even machine translation systems are on the commercial base. Such facts suggest techniques of NL processing seem so far useful and practical. But the previous NL processing systems still have some serious problems to be coped with, as we will mention just below.

One of the most serious problems is that the previous NL processing systems (machines) cannot process new expressions, i.e., unknown words and unknown phrases, which are not covered by lexical dictionary and syntactic rule dictionary (grammar) incorporated in the systems in advance. These unknown words are, for example, frequently occurred in daily newspapers (Fujisaki and Kameda, 1987). Especially, in scientific papers, which are usually regarded as the main target materials for machine translation, more unknown words and phrases are also frequently used than expected, due to expressing and conveying new ideas affirmatively.

Therefore, it is essential for NL processing systems to be able to acquire unknown linguistic knowledge (unknown words and a set of unknown syntactic rules,) so that NL processing systems become really useful for our society. From this point of view, we have been studying how to acquire the linguistic knowledge by machines (Kameda, 1992 and 1996).

Some researchers also studied unknown word processing methods (Yoshimura et al., 1989). Ishikawa et al. proposed a parsing method to analyze unknown words themselves in sentences by using a finite automaton (Ishikawa, Ito and Makino, 1993). This is a method only to detect unknown words in sentences. Kai also proposed a method in terms of speech recognition(Kai, 1996) to detect and remove just an unknown word in a sentence. Kamioka and Anzai proposed a parsing method with ability of acquiring unknown words when paring sentences (Kamioka and Anzai, 1988). By this method, sentences with unkown words are parsed with use of ICR (InComplete Reasoning) system (Kamioka and Anzai, 1988). This method is just a prototypical toy system for inferring

the part of speech and meaning of unknown words in sentences, in assumption that sentences are segmented in advance. Psychological unknown compound word meaning inference system was proposed by Kameda and Kojima, by which only unknown compound words consisting of two kanji-characters out of at most one kanji-character meaning animals. This system is interesting in terms of cognitive psychology, but just a toy system (Kameda, Hatano and Kojima, 1994) and (Kameda and Kojima, 1996).

In this paper, we propose a real-world NL processing system with abilities of acquiring unknown linguistic knowledge, especially, of both unknown words (vocabulary) and syntactic rules (grammar) by a rule-based method and by an ILP (Inductive Logic Programming) method, respectively.

Japanese unknown expressions are at first defined and Japanese unknown words are classified into three types from the NL processing point of view. Then acquisition model is proposed to describe principles of unknown linguistic knowledge acquisition briefly. A prototype system of unknown linguistic knowledge acquisition was implemented in prolog with use of progol system(Muggllton,1995) on a personal computer. Experiments were also made to evaluate the validity of the system with some amounts of text data resources.

2. Definition and Classification of Japanese Unknown Expressions

As we described before, NL processing systems cannot process all linguistically feasible expressions given to the systems as input data, firstly because any language changes from time to time in terms of vocabulary and grammar, and secondly because the systems cannot have complete knowledge of vocabulary and grammar in advance. From this point of view, we will define unknown expressions at first.

2-1. Definition of Unknown Expressions

In this paper, unknown expressions are defined as follows;

***Unknown Expressions*:** Linguistic expressions (words and phrases,) which are not produced/accepted with use of vocabulary and grammar that NL processing systems have in advance.

This definition means that unknown expressions are roughly grouped into two, according to the causes of their occurrences:

(1) unknown words
(2) unknown syntactic rules

Unknown words are also grouped into two:

(1-1) linguistically newly coined words, which are essentially not registered in system dictionary,
(1-2) words used in daily life, but not registered just in system dictionary occasionally.

Unknown syntactic rules are also grouped in the same two categories as described above like these:

(2-1) linguistically newly coined set of syntactic rules, which are essentially not registered in system dictionary,
(2-2) a set of syntactic rules used in daily life, but not registered just in system dictionary occasionally.

2-2. Definition and Classification of Unknown Words

We human can quite often understand the meaning of the first-sight unknown words easily and even unconsciously. This means it is difficult to define unknown words for human precisely. On the other hand, as we described above, NL processing systems (machines) cannot appropriately process unknown words, which are not registered in a system dictionary. From this point of view, we define unknown words, in this paper, as follows;

***Unknown Words*:** words, which are not registered in system dictionary, even though they are used as word in the real-life system of a natural language.

Unknown words in Japanese defined as above are roughly classified into three types from a NL processing point of view (Kameda and Fujisaki, 1988). Thereafter, the three types are called, in this paper, type-1, type-2, and type-3.

(1) **Unknown words of type-1 (Heterogram):** In spite that words of this type are registered in a system dictionary, but the spellings of the words are not matched with any spellings of words in the system dictionary, because spellings of the words are (not misspelled but) different from those in the system dictionary. Existence of the unknown words of this type is dependant of target language. In Japanese, any words can have several spellings, because Japanese language has some groups of characters, e.g., Kanji, Katakana, Hiragana, Romaji etc.

Example: neighbour <---> neighbor,
Tokyo <---> Tokio

(2) **Unknown words of type-2 (Compound word):** Words of this type are made up of some components (morphemes), and every of which is registered in the system dictionary, but the words themselves not. In Japanese, compound words are quite often newly coined and used for daily communication.

Examples: machine learning software,
Tokyo University of Technology

(3) **Unknown words of type-3:** Both words of this

type themselves and some or all of components (morphemes) of the words are not registered in a system dictionary.

Examples: IP address, Izu island

2-3. Definition of Unknown phrases

In this paper, unknown phrases, including also unknown sentences, are defined to be linguistic expressions that cannot be processed with use of grammar of an NL processing system, because no appropriate set of syntactic rules are not registered in the production rules of the system grammar, while the phrases consist of known words only.

3. Model and Principles of Linguistic Knowledge Acquisition

3-1. Model of Acquisition

Figure 1 shows a model of linguistic knowledge acquisition we propose in this paper. At first, system runs in normal mode to process routine works without any troubles. When any troubles happen, for example, no results are produced appropriately, the system changes the processing mode on the assumptions:

(1) Lack of a part of vocabulary,
(2) Lack of a part of grammar (syntactic rules,)
(3) Invalidity of a part of vocabulary,
(4) Invalidity of a part of grammar (syntactic rules,)
(5) Invalid or inadequate input data,
(6) Input data with noise,
(7) Inferior quality of the system performance in terms of algorithms,
(8) Malfunction of the system due to misdesign.
(9) Malfunction of the system due to program bugs.

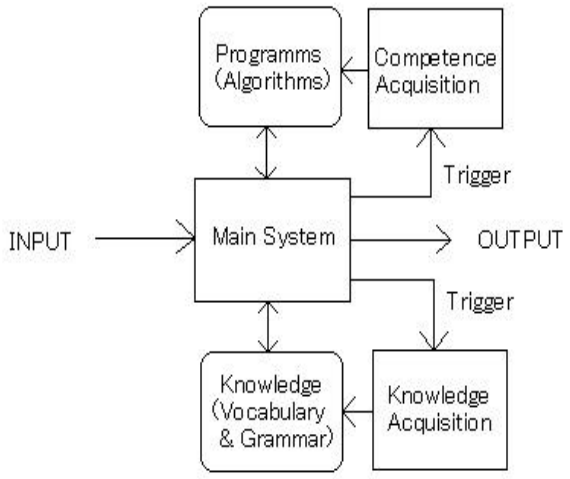

Figure 1. Acquisition model that the competence and knowledge acquisition processes are invoked by triggers when any troubles happen in main system.

Assumption (1) and (2) require the system to acquire lacking knowledge. Assumption (3) and (4) require the system to remove invalidity from the knowledge. Assumption (5) requires the system to ask valid and adequate data as input. Assumption (6) requires the system to filter out noises in data. Assumption (7) and (8) require the system to acquire or reform programs or algorithms evolutionarily. Moreover, assumption (9) requires the system to debug programs autonomously. In this paper, hereafter, we focus exclusively on knowledge acquisition in the assumption (1) and (2).

3-2. Principle of Unknown Word Acquisition

Principle of unknown word acquisition is as follows:

(1) system runs in normal mode.
(2) Trigger is pulled, when syntactic analysis failed.
(3) System changes processing mode into the one that some unknown words are in input sentence.
(4) System re-parses the input sentence on assumption that unknown words exist in the sentence, while incorporating some programs into the system to change the system mode. More details of these processes are described in (Kameda, Sakurai and Kubomura 1999).

3-3. Principle of Unknown Phrases Acquisition (Grammar Acquisition)

Grammar (a set of unknown syntactic rules) is acquired by ILP method with use of free ILP system Progol of S. Muggleton (Muggleton, 1995). More details are described in (Kamada, Kubomura and Oomori, 2000).

(1) System runs in normal mode at first.
(2) System finds and collects input sentences, which cannot be processed in both normal and unknown word acquisition mode (**3-2**).
(3) The sentences collected above are input to the Progol system, after they are compiled to be changed into Progol input files.
(4) Progol system induces a set of syntactic rules.
(5) The rules are incorporated into system grammar for normal mode process.

4. Algorithms of Unknown Word Acquisition

This chapter presents unknown word acquisition algorithms we proposed in the (Kameda, Sakurai and Kubomura, 1999) in some details.

At first, we present an algorithm for unknown word detection in the subsection 4-1. Then we also present three algorithms specifically applied to three types of unknown words in 4-2, 4-3 and 4-4, respectively. At last, we present a main algorithm for unknown word acquisition in 4-5.

4-1. Algorithm of Detecting Unknown Words in a Sentence

[1] get a part of sentence from input document and set it as word candidate.
[2] search the word dictionary for the word.
  [2-1] if the word is in it, unknown word detection is temporarily suspended to continue sentence analysis.
    [2-1-1] if sentence analysis is over consistently, it is concluded that there are no unknown words in the sentence.
    [2-1-2] if not, jump to the step [3] to resume the process suspended in [2-1] above.
[3] regard the word candidate as unknown word candidate.
[4] analyze the inner structure of the word.
[5] according to the results of the step [4], unknown word processes are invoked.
  (details will be described later)
[6] if unknown word processes fail, because the unknown word candidate is not known word nor unknown word, go to the step [2] after getting a new part of sentence as word candidate.
[7] check to see if any conflict occurs or not.
  [7-1] if no conflict occurs with the results of unknown word acquisition processes, sentence analysis and unknown word acquisition are over successfully.
  [7-2] if any conflict occurs, see if another part of speech candidate is available.
    [7-2-1] if available, go to the step [4].
    [7-2-2] otherwise, go to the step [2], after getting a new part of sentence as word candidate.

4-2. Acquisition Algorithm for Unknown Words of Type-1

[1] get a part of sentence as unknown word candidate of type-1.
[2] if the candidate doesn't consist of only Katakana character, the process is over to conclude that the candidate is not of type-1. (An assumption is so far adopted that unknown words consist of only Katakana characters. Other types are in more details described in (Kubomura, Sakurai and Kameda, 1996))
[3] substring of the candidate out of string of Katakana characters is substituted with another string with use of rewriting rules. The number of the rewriting rules is now 248 (Kubomura and Kameda, 2000).
[4] after substitution, the word with string substituted by rules are searched the word dictionary.
[5] if matched, the word is regarded as unknown word of type-1.
[6] if not, the process is over.

4-3. Acquisition Algorithm for Unknown Words of Type-2

According to surveys of Japanese texts in our laboratory, the meaning of Japanese compound words are in general expressed by the last (tail) component of words. Japanese compound words have also the structures similar to those of sentences. These facts lead us to the following procedures to process unknown words.

[1] get an unknown word candidate of type-2 as input.
[2] divide the candidate into two parts.
  (An assumption is so far adopted that any unknown compound words have always two components.)
[3] search the word dictionary for the compound word.
  [3-1] if the word is not in it, go to [4].
  [3-2] if the word is in it, display the compound word with its surface structure, deep structure, and meaning to ask users or supervisors if it is ok.
    [3-2-1] if it is not ok in [3-2], go to the step [3] again to see if there remains other word to be matched in it.
    [3-2-1'] if there remain no words to be matched in the word dictionary at all, go to the step [4].
    [3-2-2] if it is ok in [3-2], process of the compound word is over.
[4] get both information of the first and second components of the word from the component dictionary. The information is necessarily extracted from the word dictionary, because the compound word is one of type-2.
[5] construct relations between the two components within the constraints of the rules of lexical grammar described in the morpheme dictionary.
[6] if no appropriate relations are constructed with the lexical grammar, the input compound word is not appropriate, i.e., that is not the word of type-2, because of the definition.
[7] display the compound word with a candidate of its surface structure, deep structure, and meaning inferred by the system one by one, and ask the user whether the system put it onto the word dictionary for later use.
[8] the system just quits.

As you could see, what can be done by this method is to guess and compose all candidates of surface structure, deep structure, and meaning of the compound words of two components. In order to select appropriate one in the given text, outer-word context should be also used at the same time. This is the next step of this study.

4-4. Acquisition Algorithm for Unknown Words of Type-3

[1] get a part of sentence as unknown word candidate of type-3.
[2] unconditionally, input word is regarded as unknown word candidate of type-3, if no conflict occurs when sentence analysis.
[3] if any conflicts occurs, the input word is regarded as no unknown word of type-3, and then the process is over.

4-5. Integrated Main Algorithm of Unknown Word Acquisition

[1] get a sentence.
[2] pick up one of words, which have not yet been analyzed so far, sequentially from the beginning to the end of the sentence, if any.
[3] infer the part of speech of the word with use of syntactic rules.
[4] cut a word candidate from the beginning of the sentence.
[5] see if the word candidate is registered in the word dictionary.
  [5-1] if registered, get the rest part of the sentence as input.
    [5-1-1] see if the unknown word acquisition process is over.
      [5-1-1-1] if over, register the unknown word found so far into the word dictionary, and output the analysis result.
      [5-1-1-2] if not over, go to the step [2] again.
  [5-2] if not registered, go to the step [6].
[6] invoke a process for unknown word acquisition of type-2.
  [6-1] if it succeeds, go to the step [5-1].
  [6-2] if not, go to the next step [7].
[7] invoke a process for unknown word acquisition of type-1.
  [7-1] if it succeeds, go to the step [5-1].
  [7-2] if not, go to the next step [8].
[8] invoke a process for unknown word acquisition of type-3.
  [8-1] if it succeeds, go to the step [5-1].
  [8-2] if not, go to the next step [4].
[9] check to see if there are still words unanalyzed in the sentence.
    [9-1] if all words are analyzed, the process of known word detection is over.
    [9-2] if not all words are analyzed, go to the step [2].

5. Overview of Unknown Linguistic Knowledge Acquisition system and its Prototypical Implementation

5-1. Overview of the System

The figure 2 shows overview of the unknown linguistic knowledge acquisition system we propose in this paper. This system is constructed by one main system (parser) and two loosely-coupled subsystems: unknown word acquisition subsystem (UWAS) and unknown syntactic rule acquisition subsystem (USRAS).

The block of the Parser in the figure 2 corresponds to the block of the Main System in figure 1. The block of the Vocabulary and Grammar corresponds to the block of the Knowledge, and both UWAS and USRAS to the block of the Knowledge Acquisition in the figure 1, respectively. The two blocks of the Validity Checks in the figure 2, which are not depicted in the figure 1, are auxiliary processes of collecting some amounts of data to be used to prepare background knowledge for acquiring unknown syntactic rules in USRAS in the figure 2.

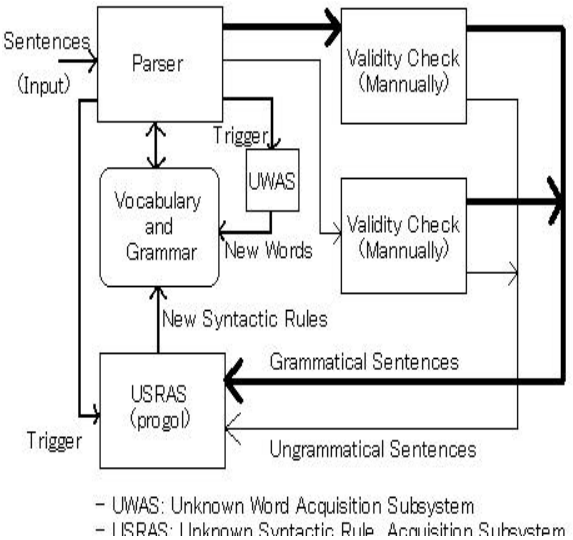

Figure 2. Overview of Unknown Linguistic Knowledge Acquisition System

5-2. Implementation of Prototype System

(1) **UWAS (Unknown Word Acquisition Subsystem):**

The part of UWAS with the algorithms described above was implemented on a note-type personal computer (FMV-BIBLO NU, Fujitsu, 96MB RAM, 2.1GB HDD). As operating system (OS), Microsoft Windows95 was adopted to run the system. Program was written in programming language Prolog (Arity/Prolog Ver.5.1). The system consists of 8 modules: main module, text input module, parsing module, three unknown word processing modules for all three types of unknown words respectively, word registration module, and auxiliary functions module. The size of the program (except dictionaries with about 100,000 words) is about 10,000 lines. Moreover, component dictionary, concept dictionary, word dictionary, morpheme word rule dictionary and

syntactic rule dictionary were also implemented in Prolog as knowledge base for the system.

(2) **USRAS (Unknown Syntactic Rule Acquisition Subsystem):**

The part of USRAS was implemented on the base of an ILP system "Progol" with three utilities, which were written in C programming language in order to prepare for input files to Progol system by compiling unparsed sentences because syntactic rules for the sentences are not registered in the system knowledge-base (Kameda, Kubomura and Oomori, 2000).

6. Evaluation of the System

6-1. Experiments for System Evaluation

In order to examine the validity of the method we adopted, the system was evaluated with some amounts of text resources.

Experiment No.1

(1) **Objective:** To clarify the performance and efficiency of the unknown word acquisition of type-1 with use of algorithms in the subsections 4-1 and 4-2.

(2) **Materials:** Dictionary("Lexicon of new katakana words," Astro-Education System, 1994), newspaper() articles on the first pages in the Asahi and the Nikkei newspapers), and books(novels and essays) are used to collect words of Katakana string.

(3) **Method:** Check all words, which come out of the system, when words of Katakana string are input to the system as evaluation material.

(4) **Results:** 121 input words (96.8%) were correctly processed out of 125 input words. Correct answers were generated by the system as the 3rd or 4th resolution in average (Kubomura, Sakurai and Kameda, 1996), (Kubomura, Kameda and Miyabayashi, 1998), (Kubomura and Kameda, 2000).

Experiment No.2

(1) **Objective:** To estimate the performance of the unknown word acquisition of type-2 with us of algorithms in the subsections 4-1 and 4-3.

(2) **Materials:** 501 unknown words (188 lexemes) of type-2 in the newspaper articles of amounting up to 84 days from one of the most famous and wide-spread daily newspaper (Asahi newspaper of 1982), (Fujisaki and Kameda, 1987).

(3) **Method:** Check the outputs manually of the process of the unknown words of type-2 by the type-2 module.

(4) **Results:** 383 words (76.4%), 138 lexemes (73.4%) were correctly processed out of 501 words, 188 lexemes (Kubomura, Sakurai and Kameda, 1996).

Experiment No.3

(1) **Objective:** To estimate the performance of the unknown word acquisition of type-3 with use of algorithms in the subsections 4-4 and 4-5.

(2) **Materials:** 56 simple sentences with unknown words of class-3 collected from a textbook "Spanish basic words 2000."

(3) **Method:** Check how many sentences are processed correctly, out of 56 sentences with unknown words of type-3.

(4) **Results:** 54 sentences with an unknown word of type-3 were correctly processed (Kubomura, Sakurai and Kameda, 1996).

Experiment No.4

(1) **Objective:** To confirm the feasibility and validity of the ILP method for syntactic rule acquisition, with use of the Muggleton's Progol system (Kameda, Kubomura and Oomori, 2000).

(2) **Materials:** 422 sentences in a English textbook of senior high schools (Simaoka et al. "Sunshine English Course 3 (4-th Edition), Kairyudo (1999)") with 884 lexemes, 3,489 occurrences of words, and 1247 words in terms of part of speech in the text.

(3) **Method:** Induce a set of unknown syntactic rules, if possible, with use of Progol system, and then check running time and quality of induced syntactic rules manually.

(4) **Results:** 7 syntactic rules were induced by Progol system automatically. 4 out of 7 were linguistically valid. Running time was about 28 hours 15 minutes on a personal computer of WINDOWS95 OS with the clock frequency of CPU 400MH and the main memory 64MB.

6-2. Considerations of Evaluation Experiments

According to the evaluation experiments said above and some other auxiliary experiments not mentioned in this paper, our system has the following characteristics:

(a) the system has fundamental capabilities to acquire unknown words in Japanese written-language document.

(b) the system can automatically improve its performance in terms of processing time.

(c) the methods proposed are expected to be partially applied to NL processing systems for languages other than Japanese.

(d) the firstly-coming-out-resolutions are not always the appropriate ones in this system.

(e) for this reason, interaction between users and the system is important, so that the system will be in practical use.

(f) incomplete knowledge, i.e., incorrect words are likely to be acquired, if the system runs with no interaction with users. (This means some

mechanisms of consistency resolution are necessary for the system to run fully automatically).

(g) the efficiency of the system is now not good enough to be incorporated into practical application software.

7. Conclusions

In this paper, we proposed an unknown linguistic acquisition system based on a rule-based method. Japanese unknown words were at first defined and classified into three classes for NL processing systems. Model and principles of acquisition were also described. Acquisition algorithms for three types of unknown words were also described in some details. A prototype system was implemented in Prolog-based format on PC. Experiments were also made to evaluate the validity of the system with large amount of text data resources. As the results, fundamental validity of the system was confirmed qualitatively.

References

H. Fujisaki and H. Kameda, 1987. Vocabulary Survey of Newspaper Articles Based on Automatic Morpheme Analysis, Report of Standardization of Japanese language in Information Society, pp.61-675 (in Japanese).

H. Ishikawa, A. Ito and S. Makino, 1993. Unknown Word Processing Using Bunsetsu-automaton, 2nd class of Technical report of IEICE, LK-92-17, pp.1-8 (in Japanese).

A. Kai, 1996. A Study on Speech Recognition System for Spontaneous Speech, Doctoral thesis of Toyohachi Institute of Technology (in Japanese).

H. Kameda, 1992. A Processing Method of Class-2 Unknown Japanese Compound Words of Two Components with use of In-Word Grammar and its Prototype System Implementation, IEEE Region 10 Conference, Tencon'92, pp.710-714.

H. Kameda, 1995. Model of thought process in unknown word acquisition by machines, Technical report of IEICE, TL95-10, pp.9-16 (in Japanese).

H. Kameda and K. Kojima, 1996. Simulation of Human-like Knowledge Acquisition by Machines -Meaning inference of unknown compound words -, Evolution of Knowledge Science, Oosuga et al. (Ed.) 2-2-2, pp.23-31 (in Japanese).

H. Kameda, G. Hatano and K. Kojima, 1994. A System for Inferring Unknown Word Meaning with use of examples by analogy, The Proceedings of the 8-th Convention of Artificial Intelligence of Japan, 23-7, pp.653-656.

H. Kameda, T. Sakurai and C. Kubomura, 1999. Unknown Word Acquisition System for Japanese Written-Language Document, ACAI'99 Machine Learning and Applications, Proceedings of Workshop W01: Machine learning in human language technology, pp.62-67.

H. Kameda, C. Kubomura and N. Oomori, 2000. An Experiment of Acquiring a set of Japanese Syntactic Rules based on Inductive Logic Programming, Technical report of IEICE, TL2000-4, pp.25-32 (in Japanese).

T. Kamioka and Y. Anzai, 1988. Syntactic Analysis of Sentences with unknown words by Abduction Mechanism, Journal of Artificial Intelligence, Vol.3, pp.627-638 (in Japanese)

C. Kubomura, T. Sakurai and H. Kameda, 1996. Evaluation of Algorithms for Unknown Word Acquisition, Technical report of IEICE, TL96-6, pp.21-30 (in Japanese).

C. Kubomura, H. Kameda and Miyabayashi, 1998. Evaluation of a Speedup Method for Unknown Word Acquisition System, Proceedings of the 1998 IEICE General Conference, pp.316 (in Japanese).

C. Kubomura and H. Kameda, 2000. Evaluation of Unknown Katakana Word Processing in Unknown Word Acquisition System, Technical report of IEICE, TL99-45, pp.65-72 (in Japanese).

S. Muggleton, 1995. Inverse entailment and Progol, *New Generation Computing*, 13:245-286.

K. Yokota, H. Kameda and H. Fujisaki, 1996. A Method for Automatic Acquisition of Japanese Grammar and Unknown Cognitive Units, *Journal of Natural Language Processing*, Vol.3, No.4, pp.115-128 (in Japanese).

Yoshimura, Takeuchi, Tsuda and Shudo, 1989. Morpheme Analysis of Japanese sentences with unregistered words, *Journal of Information Processing of Japan*, Vol.30, No.3 (in Japanese).