Challenges for General-purpose Semantic Analysis Technologies

Masaaki Nagata†*, Sanae Fujita, and Hirotoshi Taira*

Abstract

We explain the language resources developed by NTT, including Nihongo GoiTaikei, one of the largest Japanese thesauri, and Lexeed, a Japanese semantic lexicon. We also explain our efforts toward developing general-purpose semantic analysis software for analyzing the meaning of words and sentences in Japanese text.

1. Toward a computer that understands human languages

To make a computer that can handle languages as we humans can do, we are building language databases with various semantic annotations and developing semantic analysis software using these databases. In this article, we first describe the language databases we have built so far including Nihongo GoiTaikei, one of the largest Japanese thesauri, Lexeed, a semantic lexicon describing the most familiar Japanese words, and Hinoki, a Japanese treebank* with various syntactic and semantic annotations. We then describe our recent work on the semantic analysis of words and sentences using these databases, namely word sense disambiguation and predicate argument structure analysis. We also describe our efforts toward textual entailment recognition, which has recently been attracting a lot of interest in the research community as *middleware* for semantic analysis to build advanced language processing applications such as question answering and summarization.

2. Nihongo GoiTaikei

Nihongo GoiTaikei is a Japanese thesaurus that defines word senses of about 400,000 words using about 3000 semantic categories [1]. It defines three different hierarchies of semantic category for common nouns, proper nouns, and verbs, in which the common noun category is most frequently used.

Part of the common noun semantic category hierarchy is shown in **Fig. 1**. The hierarchy is defined using is-a relations and part-of relations. For identification, each category has a name and a number starting from 1. For example, the Japanese word *raitaa*, which is derived from two English words "writer" and "lighter" transliterated into the same Japanese string, is associated with two different semantic categories, "353: Author" and "915: Household appliance". By following the is-a link, we can learn that the former can be an agent (3: Agent) while latter is a physical object (533: Physical object).

As Nihongo GoiTaikei was originally developed for describing syntactic pattern conversion rules in the Japanese-to-English machine translation system ALT-J/E, it is useful for pattern-based text mining tasks. For example, suppose you try to find the places where people go shopping by extracting Japanese phrases that match the pattern "X-*de kau*" (buy at X), where X represents a variable. You can easily give semantic constraints to the noun phrases to be extracted by constraining the semantic category of the matching nouns to "388: Place".

† NTT Communication Science Laboratories
Souraku-gun, 619-0237 Japan
Email: nagata.masaaki@lab.ntt.co.jp

* Treebank: A treebank, by analogy to a databank, is a large collection of syntactic trees.

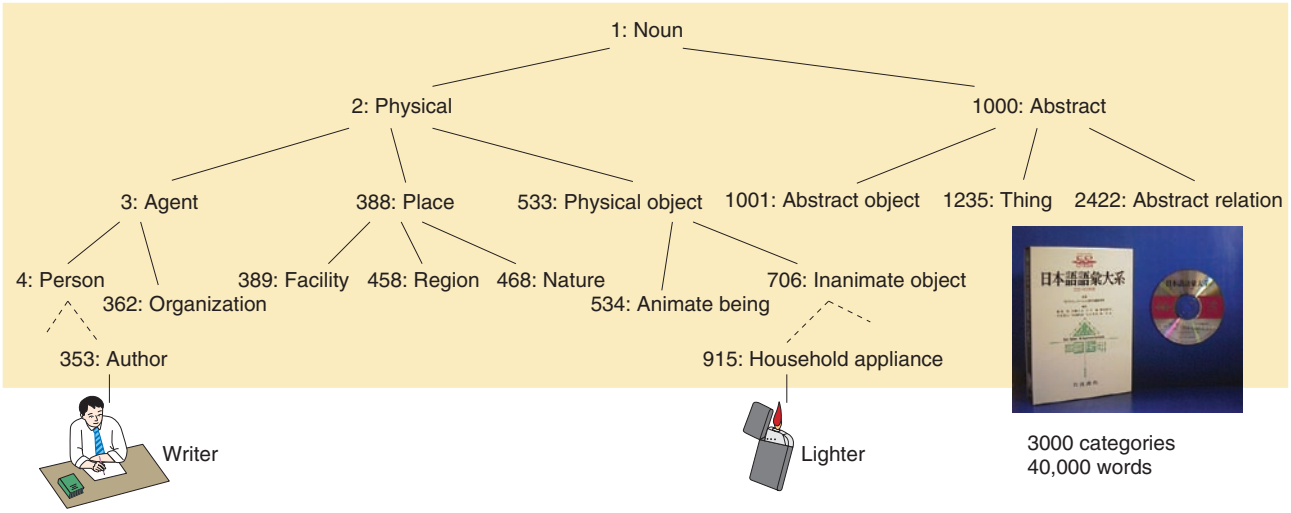

Fig. 1. Common noun semantic categories of Nihongo GoiTaikei.

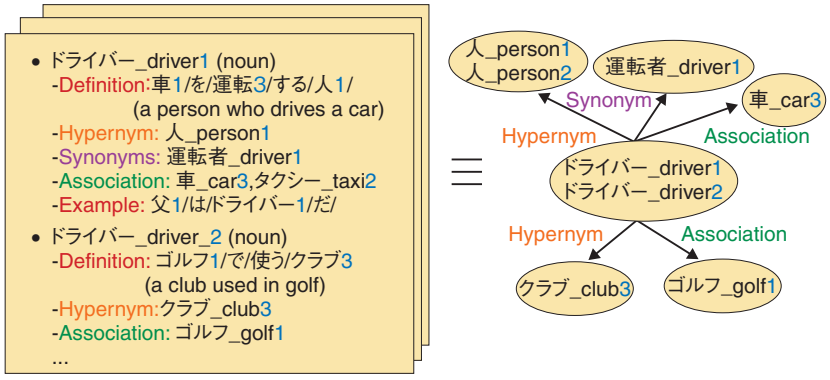

Fig. 2. Japanese semantic lexicon Lexeed.

3. Japanese semantic lexicon Lexeed and Japanese treebank Hinoki

Lexeed [2] is an electronic dictionary in which basic Japanese words (28,000) and word senses (46,000) are selected by using a psychological measure, familiarity [3], which represents the degree to which the average person is familiar with that word. Relations between word senses such as hypernym, hyponym, and synonym are systematically labeled and each sense is associated with the categories in Nihongo GoiTaikei.

The description of *doraiba* (driver) in Lexeed is shown in **Fig. 2**. While Nihongo GoiTaikei only has two semantic categories "292: Driver" and "942: Machine tool", Lexeed has a definition and an example for each sense. For the first word sense, the definition is "a person who drives a car" and the example is "My father is a driver."

In Lexeed, the definition is written in only basic words, and each word in the definition is associated with its word sense number in Lexeed. In English, the Longman Dictionary of Contemporary English is well known for having descriptions that are written using only basic vocabulary. As far as we know, Lexeed is the only such self-contained dictionary for Japanese.

We are building language resources, not only for word meaning but also for sentence meaning. Hinoki [4] is a Japanese treebank that has syntactic and semantic annotations for about 200,000 sentences or phrases including the definitions and examples in Lexeed and excerpts from newspapers. It defines the syntactic and semantic structure of a sentence/phrase based on a language theory called HPSG (head-driven phrase structure grammar). An example of the

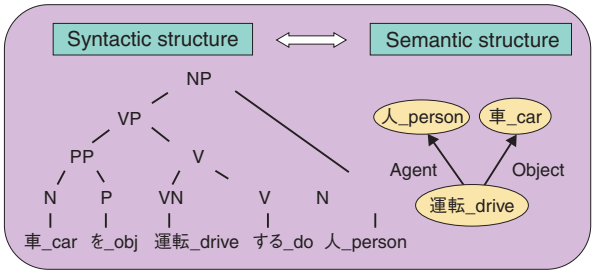

Fig. 3. Japanese treebank Hinoki.

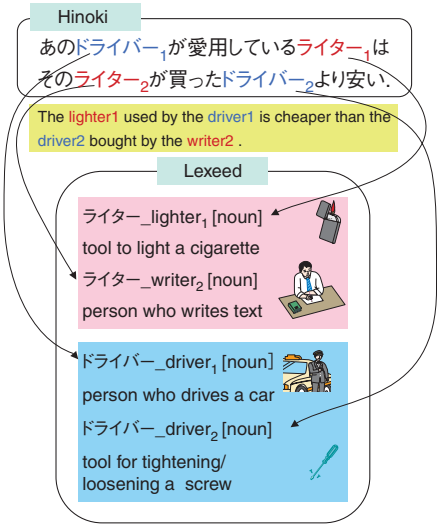

Fig. 4. Word sense disambiguation.

syntactic structure and semantic structure of a Japanese phrase meaning "a person who drives a car" is shown in **Fig. 3**. Syntactic structure represents how smaller syntactic constituents such as N (noun) and V (verb) recursively form a larger syntactic constituent such as NP (noun phrase) and VP (verb phrase). Semantic structure represents how an event described by the expression is formed from semantic elements such as agents, objects, and predicates.

As the design of Lexeed is based on psychological insights, it is useful for educational applications and user interface design, in which we have to measure the psychological burden that the text imposes on the user. As Hinoki forms a large-scale language dataset with unified syntactic and semantic annotation from the word level to the sentence level, it provides us with an ideal research playground for semantic analysis.

4. Word sense disambiguation

We are developing a set of general-purpose semantic analysis software for Japanese text using the language database described above. First, for word meaning analysis, we are studying word sense disambiguation. A large number of words have more than one word sense, and the word sense of a particular usage can be determined only by its context. For example, "driver" has several meanings such as person who drives a car, golf equipment, screwdriver, and software for driving something. Word sense disambiguation is the task of selecting the appropriate word sense in a given context from a list of word senses defined in the dictionary.

An example of a sentence taken from Hinoki, where the sense for each word is labeled using the word senses defined in Lexeed, is shown in **Fig. 4**. Using such manually labeled training data of about 200,000 sentences, we trained a sequential classifier, which determines the sense of each word from the features of its context words such as its surface form and part of speech. We call this software a *word sense tagger* [5].

As the usage of words is so diverse, even if we restrict the number of target word senses to be discriminated to the ones defined in Lexeed (about 50,000), we only have a few training examples for each word sense. We therefore devised a method of improving the disambiguation accuracy by dividing the process into two steps: we first select a coarse-grained category using the semantic hierarchy of Nihongo GoiTaikei and then select a fined-grained word sense based on the category [6].

Word sense disambiguation can be used for word sense-based information retrieval. For example, suppose "driver" is entered as a query term. If the system can tell two senses of this word in advance, it can display the documents about cars and golf separately as retrieval results.

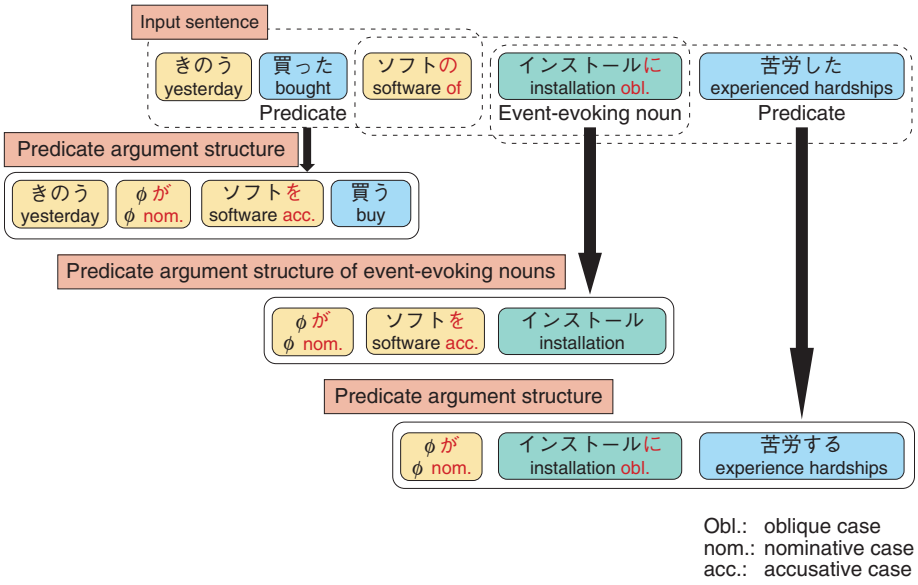

Fig. 5. Predicate argument structure analysis.

5. Predicate argument structure analysis

As a first step for sentence meaning analysis, we are studying predicate argument structure analysis. A predicate such as a verb and an adjective is the central element of a sentence, and it describes the movement or state of an event. The argument is a person or thing associated with the event. In Japanese, an argument is expressed by a noun followed by a case particle. Predicate argument structure analysis is the task of identifying and associating the arguments relevant to each predicate.

An example of predicate argument structure analysis, in which three predicate argument structures are extracted from a sentence, is shown in **Fig. 5**. Note that, not only verbs such as "buy" and "experience" but also a noun "installation" can represent an event. These nouns are called event-evoking nouns.

We devised a predicate argument structure analysis method based on structure learning, where the analysis is regarded as a structure transformation from a word dependency structure to a predicate argument structure [7]. The model for performing the task is trained from manually annotated data.

Predicate argument structure analysis can be used for event-based information retrieval. Current search engines are not good at handling how-to questions. By using a predicate argument structure, we can handle various expressions that refer to the same event such as "purchase software", "purchased software", "purchase of software", and "software purchase". The structured representation gives greater benefit as the described event gets complicated.

6. Text entailment recognition

How can we evaluate a computer's degree of understanding of human language? Text entailment recognition, a task recently proposed in the research community, might answer this question [8]. When two texts, namely "text" and "hypothesis", are given to the system, the task of text entailment recognition is to determine whether the text entails the hypothesis. The task is almost as the same as reading comprehensions tests for humans such as "read the following text and answer whether the following statements are correct or wrong".

An example of a text entailment recognition task is shown in **Fig. 6**. In order to determine whether the text entails each hypothesis, the system must analyze the predicate argument structure in order to determine whether the events expressed by the two underlined phrases are the same. Some lexical knowledge is required to tell whether the two dotted-underlined phrases are paraphrases of each other. It is also necessary to perform logical inference to determine that if the amount of bioethanol production in the USA is the highest in the world, then the USA produces bioethanol.

If we can implement the text entailment recognition

module, it can be used as a general-purpose building block for advanced text processing applications such as question answering and summarization. How text entailment recognition will be used in question answering is illustrated in **Fig. 7**. The question is converted into a declarative sentence and if it can be entailed by some text in the database, the answer will be the argument in the text that matches the question word.

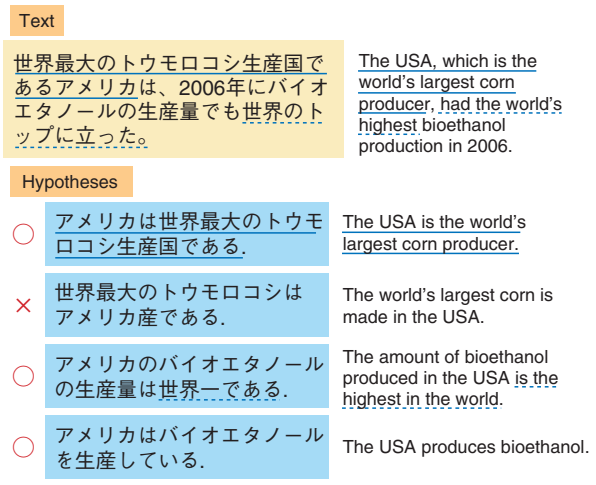

Fig. 6. Text entailment recognition.

Research on text entailment recognition for Japanese has just begun. We started to work on it in 2007 and have made about 200 benchmark data so far. We hope to report on our results in the near future

References

[1] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi, "Nihongo GoiTaikei [Japanese Lexicon]," Iwanami Shoten, 1997 (in Japanese).

[2] K. Kasahara, H. Sato, F. Bond, T. Tanaka, S. Fujita, T. Kanasugi, and S. Amano, "Construction of a Japanese Semantic Lexicon: Lexeed," IPSJ SIG Technical Reports, Vol. 2004, No. 1 (NL-159), pp. 75–82, 2004 (in Japanese).

[3] S. Amano and T. Kondo, "Nihongo-no Goitokusei [Lexical Properties of Japanese]," Sanseido, 1999 (in Japanese).

[4] F. Bond, S. Fujita, and T. Tanaka, "The Hinoki Syntactic and Semantic Treebank of Japanese," Language Resources and Evaluation, Springer Netherlands, DOI: 10.1007/s10579-007-9036-6, 2006.

[5] T. Tanaka, F. Bond, T. Baldwin, S. Fujita, and C. Hashimoto, "Word Sense Disambiguation Incorporating Lexical and Structural Semantic Information," Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 477–485, Prague, June 2007.

[6] S. Fujita, F. Bond, and A. Fujino, "Word Sense Disambiguation using Superordinate Semantic Classes," Annual Meeting of the Association for Natural Language Processing, pp. 568–571, 2008 (in Japanese).

[7] H. Taira and M. Nagata, "Predicate-argument Structure Analysis Using Structure-based Learning," Annual Meeting of the Association for Natural Language Processing, pp. 556–559, 2008 (in Japanese).

[8] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan, "The Third PASCAL Recognizing Textual Entailment Challenge," Proc. of the Workshop on Textual Entailment and Paraphrasing, pp. 1–9, Prague, June 2007.

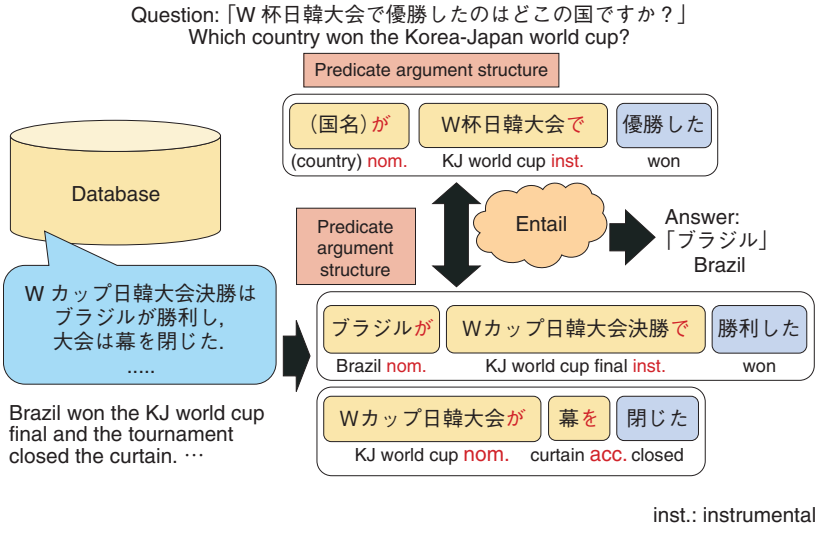

Fig. 7. Question answering using text entailment recognition.

Masaaki Nagata
Senior Research Scientist, Supervisor, Group Leader, Natural Language Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received the B.E., M.E., and Ph.D. degrees in information science from Kyoto University, Kyoto, in 1985, 1987, and 1999, respectively. He joined NTT in 1987. He was with ATR Interpreting Telephony Research Laboratories from 1989 to 1993. His research interests include natural language processing, especially morphological analysis, named entity recognition, parsing, and machine translation. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan, the Information Processing Society of Japan (IPSJ), the Japanese Society for Artificial Intelligence, the Association for Natural Language Processing (NLP), and the Association for Computational Linguistics (ACL).

Sanae Fujita
Research Scientist, Natural Language Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

She received the B.E. degree in engineering from Osaka Prefecture University, Osaka, and the M.E. degree in engineering from Nara Institute of Science and Technology, Nara, in 1997 and 1999, respectively. She joined NTT Communication Science Laboratories in 1999. Her research interests include natural language processing, especially semantic analysis, word sense disambiguation, and knowledge acquisition. She is a member of NLP and ACL.

Hirotoshi Taira
Research Scientist, Natural Language Group, Innovative Communication Laboratory, NTT Communication Science Laboratories.

He received the B.S. degree in chemistry and the M.S. degree in solid-state chemistry from the University of Tokyo, Tokyo, and the Ph.D. degree in information science from Nara Institute of Science and Technology, Nara, in 1994, 1996, and 2002, respectively. He joined NTT in 1996. He was with the Business Intelligence Deployment Center, Research and Development Headquarters, NTT DATA, from 2005 to 2007. His research interests include natural language processing and machine learning and bioinformatics. He is a member of IPSJ and NLP.