English-Japanese Dictionary System with Intelligence

Ryo Nagaino Yuji Sagawa Noboru Sugie
Meijo University
1-501 Shiogamaguchi, Tempaku-ku, Nagoya, Aichi, 468-8502, Japan
c3002020@meijo-u.ac.jp, sagawa@meijo-u.ac.jp, sugie@meijo-u.ac.jp
TEL:+81-52-832-1151, FAX:+81-52-832-1169

Summary: With the spread of computer network, chances of reading English texts are increasing remarkably for Japanese people. We, native speaker of Japanese, read a text in English with an aid of an English-Japanese dictionary. An electronic dictionary reduces efforts to look up word entry. However, it is not easy for beginners of English to choose correct sense of word when the word has a lot of ambiguities. They resolve the problem by spending a lot of time, or give up reading the sentence. We are developing a system, to which natural language processing technologies is applied, which assist them in looking up a word by disambiguating it. The purpose of this paper is to show outline of the system.

Keyword: electronic dictionary, word sense disambiguation, computer assisted learning (CAL)

1. Introduction

In Japan, books, software, and schools to aid English learning are increasing. With spreading of PC and network, it becomes easier to visit a website and to access a document written in English. Moreover, some of companies require English skill in order for a person to be promoted.

Thus, it is important for Japanese to be able to understand English. As a result, we, non-native speaker of English, must learn English. We use an English-Japanese dictionary when we read a text in English. Particularly, an electronic dictionary helps us to reduce the load to look up a word.

Problem left is load to choose appropriate meaning from a lot of meanings one entry has. This work needs knowledge and experience of English, so beginners seem to need some help.

The purpose of this study is to add a function to disambiguate a word sense and a function to support learning English to English-Japanese dictionary system.

2. The English-Japanese Dictionary System with Intelligence

Figure 1 shows the outline of the flow of our system. We apply techniques of word sense disambiguation developed in the field of natural language processing to our system.

A user inputs the text into the system. When he/she encounters an unknown word in the text, he/she clicks on the word to order the system to consult the word. Then, the sentence included the word is parsed by the system. If the input sentence is the same as an example in the dictionary or the word is

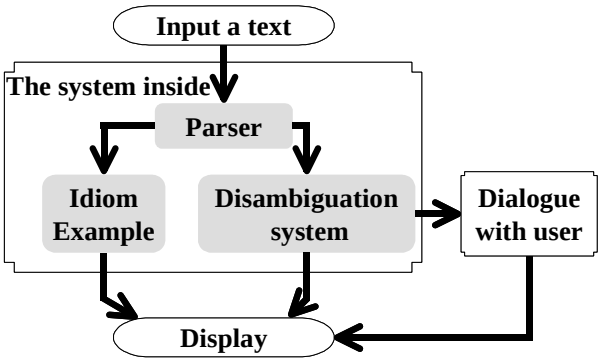


Figure1: a flow diagram of Electronic English-Japanese Dictionary System with Language Experts' Intelligence

used as a part of an idiom, the system displays them. When it was not so, the sense of the word is disambiguated by the disambiguation system and displayed. However, when the system cannot disambiguate the word, it is solved with help of the user via interaction.

The main feature of the system is "disambiguation system" in figure 1. This is why our system has "intelligence", which previous dictionary systems do not have.

Although various methods of word sense disambiguation has been developed so far [1], the prerequisite of our system differs from those of them in the following senses:

- It is enough for our system to select "proper corresponding Japanese words" in dictionary rather than specify "word sense".
- The system needs not to always limit a solution to one.
- The system should be able to explain why Japanese word is selected. This is because the system is used to learn language.

3. Disambiguating the senses of the preposition 'by'[2]

3.1 Method of disambiguation

To show how our system disambiguates word sense, we explain the case of preposition 'by' as an example.

Statistical way of word sense disambiguating, which is the mainstream of, cannot be used, because our system should be able to explain why the proper term was selected. So, we select a rule-based approach. In order to make rules, we refer to contents of " Shogakukan Random House English-Japanese Dictionary Second Edition "[3]. The number of categories of sense of 'by' is 23 in the dictionary. Moreover, in order to make rules, we used features, that a preposition in a sentence is decided by the construction of the sentence, by the verb, or by nouns at the front of preposition and/or the rear.

Rules are distributed among categories based on the following three kinds of information. The system tests which rules are applicable and proper terms to correspond are selected.

(1) Rule based on syntactic information

For example, when the main verb is modified 'by' and the sentence is the passive voice, this 'by' means "agency". Thus, even if the system does not do semantic analysis, there is the case that it can disambiguate.

(2) Rule based on related nouns

If the noun following 'by' means "measure", as "too many by one" or "win by a boat's length", that 'by' specifies "amount or degree of something". To check this, meaning of 'one' and 'length' are required.

We use is_a hierarchy.

This rule can be represented as follows;

“If the sense of the noun following ‘by’ is a subordinate concept of ‘measure’, the category of the proper term is ‘amount or degree of something’”

In the case of “too many by one”, the system checks is_a hierarchy as follows: one -> digit -> integer -> number -> definite quantity -> measure.

In the case of “win by a boat’s length”: length -> dimension -> measure. These are matched this rule.

Thus the proper sense can be selected in the system. We use WordNet[4] hierarchy to analyze sense of noun.

In addition, rules not only use the noun following ‘by’ but also use the noun followed by ‘by’.

(3) Rule based on sense of the related verb

The preposition ‘by’, as function word, has relation with the verb as well as the noun. Therefore, there is possibility to disambiguate of ‘by’ by checking the sense of verb modified by it, the preposition ‘by’.

The question here is the how to express semantic information of the verb. To dispose as simply as possible, verbs is classified into some categories. Concretely basic verb, which Schank[5] suggested, is given to each verb. However, because the original set of these verbs is not enough for our purpose, we add some new basic verbs. The number of basic verbs is 14.

One rule includes clues both the verb and the noun following ‘by’.

3.2 Example

In this section, we show how to disambiguate by using illustrative sentences, as follows.

<1> The phonograph was invented by Thomas Edison.

<2> The production of foodstuffs increased by 50 percent.

<3> Eve had two sons by Adam.

<4> He caught me by the arm.

Table 1 is syntactic information and

Table 1: Syntactic Information and Semantic in examples

	structure	the sence of the noun followed by 'by'	the sence of the noun following 'by'	basic verb
<1>	passive voice		human	MBUILD
<2>	active voice		amount	ATRANS
<3>	active voice	human	human	POSSESS
<4>	active voice	human	a part of body	GRASP

Table 2: Parts of Rules to Disambiguate preposition ‘by’

	structure of the sentence	the sence of the noun followed by 'by'	the sence of the noun following 'by'	relation between the noun followed by 'by' and following	basic verb	Proper Term
Rule 1	passive voice					～niyotte
Rule 2			amount			～dake
Rule 3		creature or thing	creature	child and parent or work and creater		～karaumareta
Rule 4		animal or thing	a part of animal or thing	the following 'by' is part of the followed	GRASP	～notokorowo

semantic in examples.

Table 2 is parts of rules to disambiguate preposition 'by'.

Compare the information on Table 1 with Table 2, we turn out that <1> correspond to Rule 1, <2> to Rule 2, <3> to Rule 3, and <4> to Rule 4.

Thus, in disambiguation system, the comparison of "syntactic information and semantic in a sentence" and "Rule" is done.

3.3 Evaluation

In this section, we show the results of our evaluation test. The following texts are used in the test.

(1)D.O.Hebb "Organization of Behavior" Chapter 4 (1961).

(2)P.H.Winston "Artificial Intelligence Third Edition" Chapter 10 (1993).

(3)D.Jensen, M.Atighetchi, R.Vincent, V.Lesser "Learning Quantitative Knowledge for Multiagent Coordination" (1999).

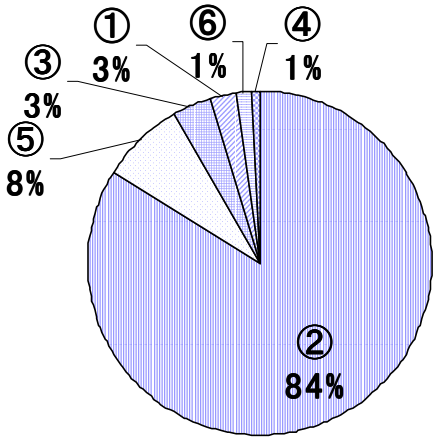
(4)G.A.Keim, N.M.Shazeer, M.L.Littman "PROVERB : The Probabilistic Cruciverbalist" (1999).

(5)J.Y.Chai, A.W.Biermann "The Use of Word Sense Disambiguation in an Information Extraction System" (1999).

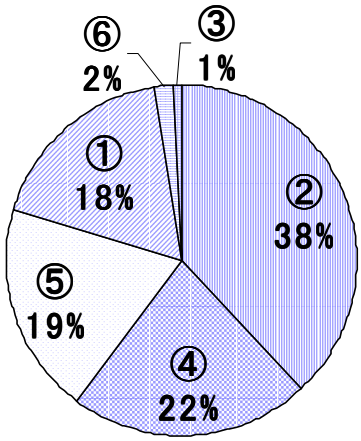
(6)M.Veloso, M.Bowling, S.Achim, K.Han, P.Stone "CMUnited-98 : A Team of Robotic Soccer Agents" (1999).

(7)S.L.Epstein "Game Playing : The Next Moves" (1999).

(8)R.J.Waller "The Bridges of Madison County" (1992).



(a) Technical papers



(b) Literary work

① Idiom or Example
② Rule based on syntactic information
③ Rule used the noun followed by 'by' and following
④ Rule based on sense of the related verb
⑤ Rule used only the noun following 'by'
⑥ Dialogue with user

Figure 2:Results of hand simulation studies

(1)-(7) are technical papers, and (8) is the literary work.

From the result, we conclude that more than 95% of the cases were successfully disambiguated. Particularly, rules made by syntactic information were used well in

technical papers (Figure2(a)), but rules in general were used in literary work (Figure2(b)).

4. Implementation of the system

In developing the system, we investigate what kinds of information should be supplied for users and are input by them.

4.1 Output

First, as concerns the output of the system, not only contents of a dictionary but also useful information in language learning should be displayed. In other words, the system should display proper terms disambiguated, the reason why it is disambiguated, syntactic information (a break of a phrase and relation of modification), synonyms, antonyms, and illustrative sentences.

The display of the reason can be acquired as knowledge for the user. Its knowledge will be used by the user when he/she encounters a similar sentence. And if the user request illustrative sentences, antonyms, and synonyms, knowledge of target word is understood by him/her more and more, and he/she increase his/her vocabulary.

4.2 Input

Second, as regards the input of the system, kinds of information that the system needs are a text, which the user will read, the information of the context, and a word that he/she wants to look up in a dictionary.

It is necessary to acquire the information of the context to disambiguate successfully, but it is hard to obtain automatically.

For reasons mentioned above, we design the flow of interaction between users and our system as follows;

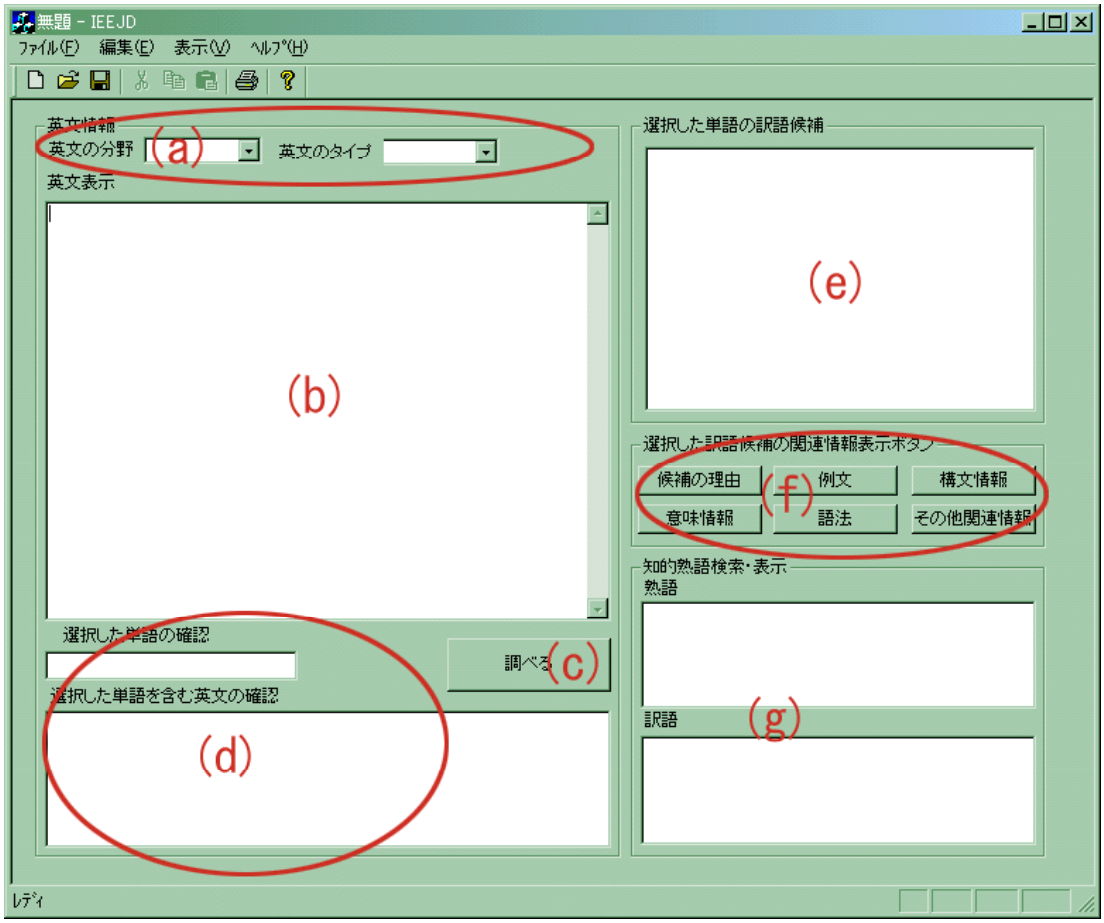
(1) The user specifies a text and a type / a realm of the text.
(2) The user specifies a word looked up in dictionary.
(3) The system gets propositional Japanese terms by disambiguating the sense or by comparing with idioms or examples and displays these.
(4) The user selects the best proper term from proposition. If there is not the best proper term in those, others corresponding Japanese term is displayed and the user select.
(5) If it is necessary, the user requests the reason and the information concerned (synonyms, antonyms, illustrative sentences) for the system.
(6) If the user needs, he/she repeats (4) and (5). Therefore, he/she can certainly select the best proper term.

When one ambiguity word is disambiguated, there is possible to disambiguate another word. So, (6) is necessary.

We are developing the system on Windows 2000 with Visual C++ 6.0, and PC-PATR [6], which is developed by SIL International as a parser.

5. Interface

The interface of our system is showed figure 3.



(a) Field to specify a type / a realm of the text
(b) Field to display the text which the user will read
(c) Button to disambiguate the object word
(d) Fields to display the word, which the user selected, and the sentence, which include the word
(e) Field to display proper terms by disambiguating
(f) Buttons to select the information that the system provides for the user in order to learn
(g) Fields to display the idiom

Figure 3: Interface of the system

In (f), when the user wants deep knowledge of the object word, he/she pushes the button corresponded to the information that he/she wants. Kinds of information are the reason, illustrative sentence, syntactic information, semantic information, usage, and the information concerned (synonyms and antonyms), as we described previous section.

In (g), if the system judges the object word to be a part of the idiom, the system displays it in the special way.

In the special way, for example, when the system discovers the idiom, “know ~ by heart”, from the sentence, “He knows the route by heart”, it displays the idiom, as “knows the route by heart”, and the proper terms, as “the route wo annki shiteiru.”

In this system, first, a user enters a text in English in (a). Next, the user reads it. When he/she encounters a word looked up in

dictionary, it is highlighted with drag. Then, he/she clicks on “look up” button, (c). And the system disambiguates, and displays proper terms in (e). After that, he/she selects the best from them. Moreover, when he/she wants to make sure of the sense of the object word and to acquire knowledge of it, he/she pushes buttons, (f), corresponded. Thus, the system provides the information concerned for him/her. If the object word is a part of the idiom, the system displays the proper terms in (e). As above, it shows the idiom and the proper terms in (g) at the same time.

6. Conclusion

In this paper, we described how to build the system disambiguated preposition ‘by’.

In the future, after the system will be completed, rules of other preposition can be made by main part of the system. We think that those rules are made by common information to obtain by entering sentences and proper terms.

Reference

[1] Hozumi Tanaka: “Natural Language Processing and Its Applications”, IEICE, pp76-80, pp213-217, (1999).(In Japanese)

[2] Ryo Nagaino, et al.: “The English-Japanese Dictionary System with Intelligence -Disambiguating the senses of the preposition ‘by’-”, Record of 1999 Tokai-Section Joint Conference of The Eight Institutes of Electrical and Related Engineers, 633(1999). (In Japanese)

[3] Tomoshichi Konishi, et al.:” Shogakukan Random House English-Japanese Dictionary Second Edition”, Shogakukan, （1994）. (In Japanese)

[4] Christiane Fellbaum: “WordNet: an electronic lexical database”, MIT Press (1999).

[5] Hozumi Tanaka, Junichi Tsujii: “Natural Language Understanding”, Ohmsha, pp.109-119, (1988). (In Japanese)

[6] Summer Institute of Linguistics: “PC-PATR -A syntactic parser-”, http://www.sil.org/pcpatr/.