COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

MAN-ASSISTED MACHINE CONSTRUCTION OF A SEMANTIC DICTIONARY FOR NATURAL LANGUAGE PROCESSING

Sho Yoshida
Department of Electronics,
Kyushu University 36,
Fukuoka 812, Japan

Hiroaki Tsurumaru
Department of Electronics,
Nagasaki University,
Nagasaki 852, Japan

Tooru Hitaka
Department of Electronics
Kyushu University,
Fukuoka 812, Japan

This is a report on the semantic dictionary for natural language processing we are constructing now. This paper explains how to obtain the semantic information for the dictionary from an ordinary Japanese language dictionary with about 60,000 items (which had already been put into machine readable form) and also explains what should be the frame for the representation of meaning of each item (word). Then a man-assisted machine procedure that embeds the semantic graph with respect to the head word of the ordinary dictionary into the frame of a head word is discussed.

1. INTRODUCTION

There are following two obstacles to the construction of a practical size semantic dictionary for natural language processing. One is that the number of words are so large that it is unknown from what and how to obtain semantic information necessary for the dictionary (tens of thousands general words should be accomodated). The other is that it is unknown how to seize and represent the meaning of each word.

We are going to settle the problems as follows. For the first problem, obtain the information necessary for the semantic dictionary from ordinary dictionary of Japanese language by analizing the sentences defining headwords. For the second problem, represent the meaning of a word by using the frame the structure of which has been studied extensively by us for a couple of decade (YOSHIDA (1982)).

Figure 1 is the outline of the construction steps of the semantic dictionary.

In this report we discuss, as preliminary stage of the construction, about the framing of the contents of the sentences defining words in the ordinary dictionary. The discussions are to be held about the followings.
(1) Features of the structure and the contents of the sentence defining (describing the meaning of) a headword.
(2) The use of semantic graph as a scheme for the representation of meaning (structure) given by the sentence that define a head word in the ordinary dictionary.
(3) The frame to represent the meaning of a word.
(4) The way of embedding the information obtained from the semantic graph in the frame given by (3).
(5) The way of obtaining the meaning (definition) of a word by unifying partial definitions (given by (4)) concerning the word.

2. INFORMATION OBTAINED FROM AN ORDINARY JAPANESE LANGUAGE DICTIONARY AND FEATURES OF THE SENTENCES THAT DEFINE HEADWORDS

We have chosen the medium size Japanese language dictionary (Shinmeikai Kokugo

jiten, Published by Sanseido) to use for constructing the semantic dictionary. The reasons for the choice of this dictionary are as follows.
(1) The number of headwords are adequate (about 60,000 items).
(2) Had already been put into machine readable form.
(3) Put emphasis on describing the meaning (signification) of a word by sentences, and avoided mere restatement in other words.
(4) Tried to classify the meanings of words into several large groups and avoided useless small groups.

Information that can be obtained from the dictionary are as follows.
(1) Syntactic information such as part of speech, inflection and conjugation.
(2) The described meaning of a headword.
(3) Several classified meanings of a multivocal words.
(4) Detailed meaning in its context and delicate shade of meaning of synonyms.
(5) Synonym, antonym and idiom (idiomatic phrase).
(6) Selection of significant words.

The features of words definitions in the dictionary are as follows.
(1) Symbols and abbreviations are used in place of sentences.
e.g. ⇨··· : make reference to ···　　五 : the five conjugations of verbs
⟷ ··· : antonym is ···　　形 : adjective
名 ··· : noun form is ···　　ド : German (language)
(2) P-N type (Noun modified by Predicate) and N-N type (Noun modified by Noun) phrase are used extensively.
(3) Words are abridged from the sentence into which the meaning is condensed, because of assuming the human use of the dictionary.
(4) The meaning of a headword is defined by viewing it from some aspects or from its upper words (or synonyms).
(5) When we look up in turn the upper words (or synonyms) of a headword, the explanation may be broken off, skipped or get into a cycle.

3. SEMANTIC GRAPH REPRESENTATION OF A WORD DEFINITION

The definition of a word should be given as accurately as possible in order to extract (meaning) information for the semantic dictionary from it. Thus the semantic graph is introduced for the representation of the word definition. The semantic graph which is a kind of semantic network consists of nodes (expressed by <···>) and arcs (or links; expressed by arrows) connecting the nodes. The semantic graph with respect to a word is the conceptual dependency graph of the sentence (called semantic graph of the sentence) defining the word. Nodes express the conceptual words (nouns, verbs, adverbs, adjectives expressed with <···>) and arcs represent the relations between them. These relations are mainly given by the functions (called link information) expressed by the relational words such as particles.

Examples of relations or link information are : 'a kind of relation', 'case relation', 'cause-effect relation', 'synonym-antonym relation', 'mode', 'attribute', 'state'. An exmaple of the semantic graph which is derived from the definition sentence "画用紙 (drawing paper) : 絵をかくためのやや厚目の紙(white and slightly thick paper for drawing a picture)" is shown in Fig.2. In Fig.2, relational words are shown explicitly in Japanese which express the link information in the semantic graph. When the more detailed relation between words is needed, the relational information expressed by the relational word is transformed into the compound relation between words such as in Fig.4. In Fig.4, a relational word "ための"(*prep.* for: used for the purpose of) is transformed into the relation which contains some words ('使用する'(use) in this case) and the relation ('object' and 'for the purpose of'). In the process of obtaining the semantic graph from a sentence, it becomes frequently necessary to supply additional information (expressed by additional words and relations) to the sentence.

4. THE FRAME REPRESENTATION OF A WORD DEFINITION

What sorts of information as a meaning of the headword are extracted from the semantic graph derived from the definition sentence in the ordinary Japanese dictionary? In the case of the semantic graph of a word '画用紙 (drawing paper)' in Fig.3, following information are extracted, by means of looking up the upper word '紙(paper)' of 'drawing paper' and the linked relations between 'paper' and the other words. ① drawing paper is a kind of paper, ② slightly thick, ③ white, ④ used for drawing a picture. These semantic information can be extracted from the semantic graph of the definitions of a word 'paper' through the views 'a kind of', 'attribute (shape)', 'attribute (color)' and 'purpose of use'. Here, 'a kind of' is a relation between 'drawing paper' and 'paper'. Furthermore, from ④, semantic information such that 'drawing paper' is the 'object (place)' for drawing a picture' is obtained. This information denotes case relationship between 'draw' and 'drawing paper'. The above can be understood, from a different point of view, that the meaning of a word 'drawing paper' is defined (described) from the views such as 'a kind of' (relation), 'case'(relation), attribute (e.g. shape, color) and 'purpose of use' (basic view point). We are aiming at constructing the frame of a word definition (called the frame of a word or the frame of definition) by means of these views. What sorts of views should be prepared for the frame of a word?

Concerning the frame we cleared that:
(i) Words (to be exact, concepts) are classified into four types, that is, concrete object, event, attribute (state) and abstract thing which are roughly correspond to concrete noun, verb, adjective/adverb and abstract word. Each word has the frame corresponding to its type.
(ii) The frame is made up of which consist of the semantic relations (such as, a kind of relationships, case relationships, cause-effect relationships and whole-parts relationships) and basic views. As basic views for 'product' thing (a kind of concrete object), we prepared ones shown in table 3.
(iii) The frame for 'event' word is mainly made up of case relationships, cause-effect relationships, a kind of relationships and views of 'mode'(e.g. tense, possibility, in progress). These relationships are shown in tables 1 and 2.

5. UNIFICATION PROCEDURE

To give the full definition of a word, the partial definitions given in the form of frames are unified into a hierarchical structure by the unification procedure. The definitions of the comparatively upper words in the hierarchical classification of the words should be given fine meaning while lower words are only given of their special meanings, because the definition inherits from the upper words. An example of the definition of a word 'drawing paper' (also, 'paper') is shown in Figs. 5 and 6, and also the definition of a word 'fire extinguishing' is shown in Fig.7. In Figs.5, 6 and 7, the semantic information connected by the dotted lines are inherited. In Fig.6, some words (e.g. 'stationary', 'instrument', 'product') are skipped between 'paper' and 'thing'. Therefore, if these words are looked up and related with 'paper', the semantic information extracted from these words should be added to the definition. There may be some shortage which can not be supplied only by these unification procedures, so that it should be given man-assistedly fine meanings to the comparatively upper words. For example, in Fig. 7, the agent of 'Remove' is 'Person' and result of 'remove' is 'vanish', therefore, also the agent of 'extinguish', 'fire extinguishing' and so on can automatically be given if we have given the information in the fine definition.

6. CONCLUDING REMARKS

Although, in our preliminary investigation, we have prospect of using the ordinary dictionary for the construction of the semantic dictionary there are many problems

to be investigated, which contain followings.
(1) Determination of the frames for abstruct noun and attribute(state).
(2) Semantic level of the semantic graph.
(3) Definitions of individual thing and event.
(4) Development of the programs and supporting programs.
(5) Definition of the technical terms.

REFERENCE:

[1] Yoshida, S. : conceptual taxonomy for natural language processing, in Kitagawa, T.(ed.), Computer Science & Technologies '82, Japan Annual Review in Electronics, Computers & Telecommunications (Ohmsha, Japan, 1982)

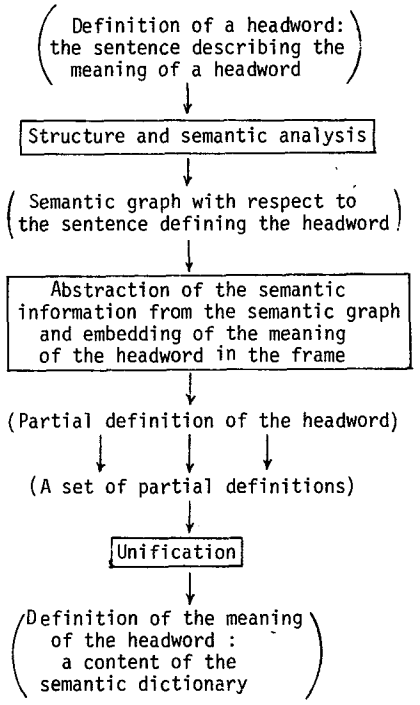

Fig.1 Steps of the construction of the semantic dictionary

Table 1 Cases used for the frame and their symbols

Case	Symbol
agent	AGT
object	OBJ
dative	DAT
source	SOU
destination	DES
insturument	INS

Table 2 Cause-effect relationship

View
means
way, method
purpose
cause
reason
result (RES)

Table 3 Basic-views necessary for defining PRODUCT

View	Basic-view
use	purpose of use(POU), (what to use)
make	material(VMA), how to make(HTM),
structure	parts, arrangement,
property	physical, mental,
time	existing time,
place	existing place,

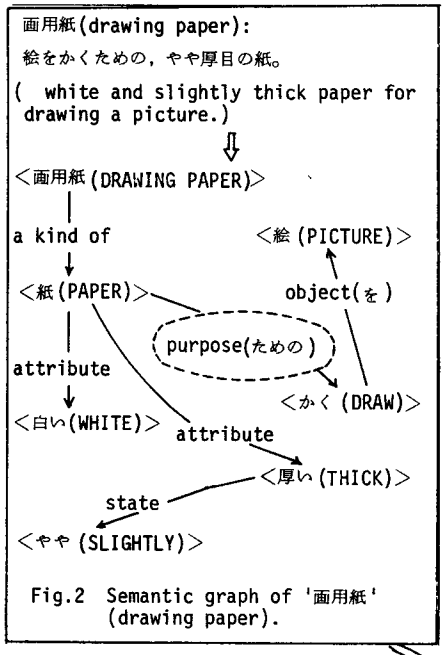

Fig.2 Semantic graph of '画用紙' (drawing paper).

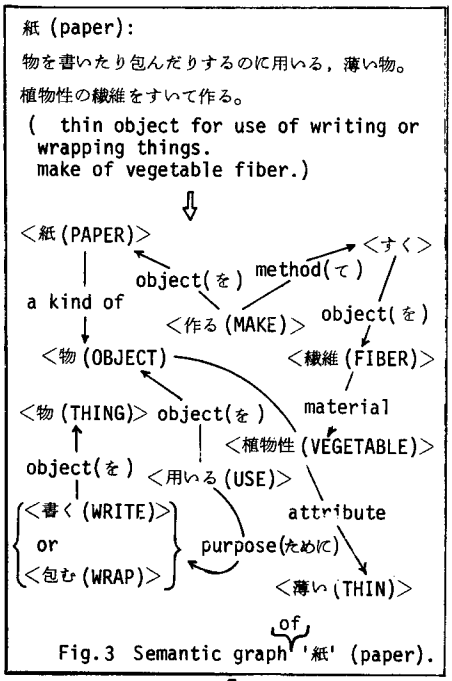

Fig.3 Semantic graph '紙' (paper).

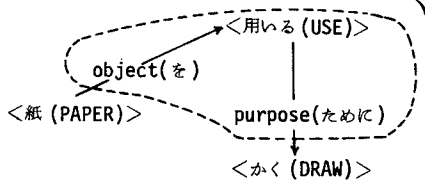

Fig.4 Detailed relation of 'ための' (use for the purpose of).

<画用紙>
(AKO (<紙>))
(POU (<かく>
(CASE (OBJ (<絵>))
(DES (*))))
(ATH (<厚い>(SDG (<やや>))))
(ACO (<白い>))
(VMA ( ))

Fig.5 Partial definition of '<画用紙>' (<drawing paper>).

<紙>
(AKO (<物$_1$>))
(POU ((<書く> (CASE (OBJ (<物$_2$>))
(DES (*))))
(<包む> (CASE (OBJ ( ))
(INS (*))))))
(HIM (<すく>(CASE (OBJ (<繊維>...
(VMA (<繊維>(VMA (<植物性>))))
(ATH (<薄い>))

Fig.6 Partial definition of '<紙>' (<paper>).

ATH : attribute(thickness)
ACO : attribute(color)
SDG : state(degree)

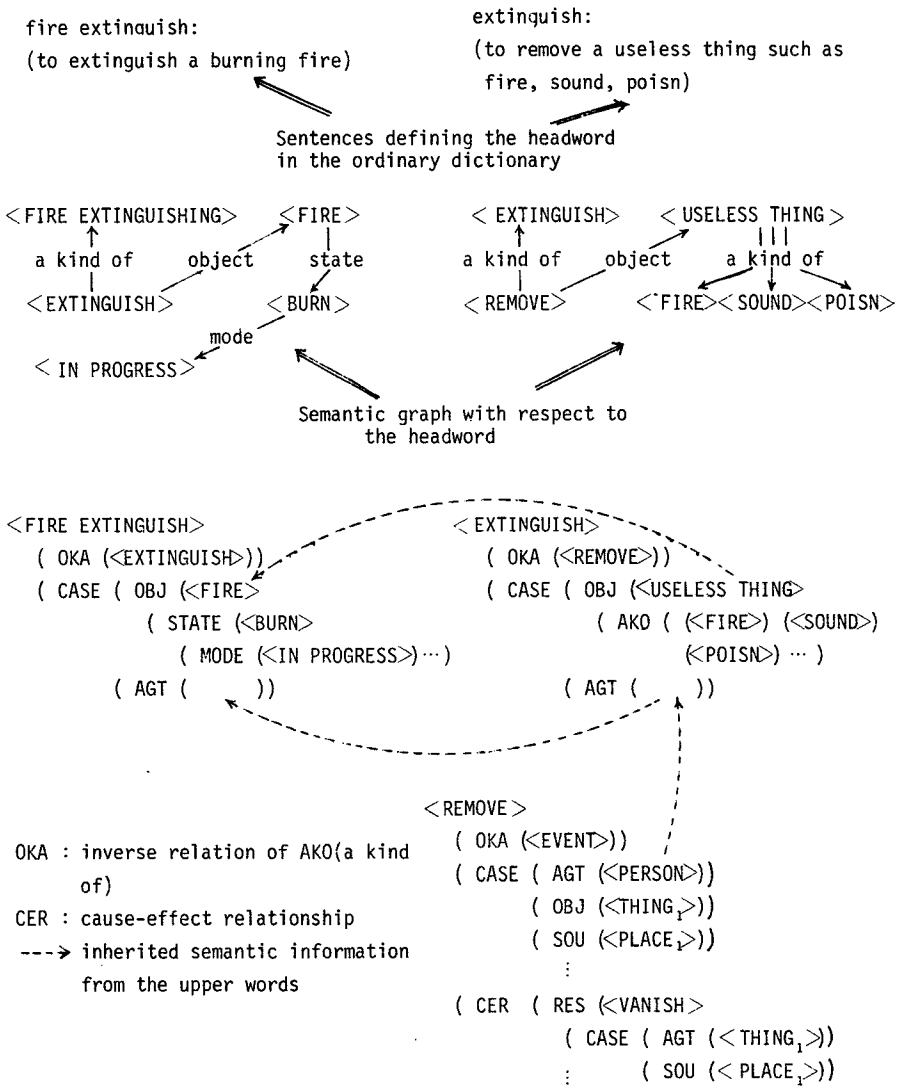

Fig.7 Partial definitions of <FIRE EXTINGUISH> and <EXTINGUISH> , and a part of the definition of <FIRE EXTINGUISHING> containing them.