AN ATTEMPT TO AUTOMATIC THESAURUS CONSTRUCTION FROM AN ORDINARY JAPANESE LANGUAGE DICTIONARY

Hiroaki Tsurumaru
Department of Electronics
Nagasaki University
Nagasaki 852, JAPAN

Toru Hitaka
Sho Yoshida
Department of Electronics
Kyushu University 36
Fukuoka 812, JAPAN

ABSTRACT

How to obtain hierarchical relations(e.g. superordinate -hyponym relation, synonym relation) is one of the most important problems for thesaurus construction. A pilot system for extracting these relations automatically from an ordinary Japanese language dictionary (Shinmeikai Kokugojiten, published by Sansei-do, in machine readable form) is given. The features of the definition sentences in the dictionary, the mechanical extraction of the hierarchical relations and the estimation of the results are discussed.

1. INTRODUCTION

A practical sized semantic dictionary (thesaurus as wide sence) is necessary for advanced natural language processing. We have been studying how to obtain semantic information for such semantic dictionary from a Japanese language dictionary(Shinmeikai Kokugojiten,published by Sanseido,in machine readable form)(1) containing about 60,000 entries.

A dictionary contains meanings and usages of practical size of general words. Especially,definition sentences{DS: a brief notation} are important sources of information for meanings of general words. Generally, DS of an entry word {EW:a brief notation} is defined by qualifying its superordinate word or synonyms or hyponyms. We call these words definition words{DW:a brief notation}.

We have been developing a system for extracting automatically DW related to EW from its DS, and for deciding the DW-EW relation(6). By this system,(hierarchical) relations among entry words in the dictionary are to be established.

We constructed a sub-system for extracting DSs corresponding to parts of speech, infrected form and meaning (definition) number of each entry word(7).

In this paper, the features of DSs in the Japanese dictionary, an outline of the pilot system and the results of experiment will be discussed.

2. FEATURES OF DS IN ORDINARY JAPANESE LANGUAGE DICTIONARY

2.1 STRUCTURE OF DS

The typical examples of DSs are as follows:

(1)【折尺(zigzag rule)】:…しまっておける(possible to be folded)ものさし(rule)。

(2)【山路(mountain path)】:「…の中の(in)小道(narrow path)」の意(the meaning of)の雅語的表現(a polite expression of)。

(3)【青蛙(blue frog)】:…大型の(large)カエル(frog)の一種(a kind of)。

(4)【嫁御(respectful daughter-in-low)】: 嫁(daughter-in-low)に対する(for)尊敬語(a respectful word)。

(5)【秋虫(autumn insect)】: スズムシ(bell-ing insect)・マツムシ(a kind of cricket)など(etc)。

Where the brackets(【…】), underline, and parentheses ((…)) denote EW, DW, and an English translation for the preceding Japanese phrase respectively.

In (1), the final word is DW and superordinate-hyponym relation(DW>EW) holds between the DW and the EW.

In (2), DW is the final word in hook brackets(「…」) and DW>EW holds. The expression "「…」の意の雅語的表現" is called a functional expression{FE:a brief notation}. The (compound) word "雅語的表現" in the FE is called a functional word{FW:a brief notation}. In this case, the FW denotes a usage of the EW.

In (3), DW is just before the FE "の一種" and DW>EW holds. In this case, the FE prescribes the DW>EW explicitly. The word "一種" is the FW.

In (4), DW is just before the FE "に対する尊敬語" and the synonymous relation(DW≡EW) holds between the DW and the EW. The FW "尊敬語" denotes a usage of the EW.

In (5), two DW<EWs hold, that is, the DW "スズムシ" < the EW "秋虫" and the DW "マツムシ" < the Ew "秋虫". In this case, the number of DWs are more than one, DW isn't modified and the FE is the word "など". The FW is identical with the FE. (Notes: "など" is a sub-postpositive signifying exemplification.)

The features of DSs in the dictionary are as follows:

(a) Generaly, the final word in DS is DW.
(b) In some cases, the final expression in DS is FE assigning semantic relation between DW and EW, and DW is just before the FE.
(c) Genraly, DW is modified by another phrase(modifier).
(d) In some cases, DS contains more than one DW.

The following general structure is obtained according to these features.

… ([MODIFIER]・DW)*・[FE]。

Notes) […] : optional constituent
(…) : required constituent
* : sequence of coordinate constituent(e.g. ・,と)
・: concatination symbol which is diferent from coordinate constituent(・)

For convenience of explanation, the general structure is divided into the following two types.

(I) TYPE I : … ([MODIFIER]・DW)*。
(II) TYPE II : … ([MODIFIER]・DW)*・FE。

In TYPE I, the final word is DW. In TYPE II, the final expression is FE, and DW is just before the FE.

2.2 DW-EW RELATION IN DS

We will propose the following assumptions according to above-mentioned features in order to extruct the DW-EW relations from DSs of the general structure.

① When DS is in TYPE I, DS≡EW. Because DS is a phrase (or a compound word) as wide sence.

② When DS is in TYPE II, SS ρ_{FE} EW.
Where ρ_{FE} is binary relation assigned by FE, and SS is the shortened DS corresponding to ([MODIFIER]・DW)*.

③ [MODIFIER]・W ≦ W

④ ([MODIFIER$_i$]・W$_i$)* ≧ [MODIFIER$_j$]・W$_j$
Where i,j=1～n, W is arbitrary word.

The following general algorithm for deciding the DW-EW relations is obtained by means of these assumptions.

(I) DS is in TYPE I (DS dosn't include FE),
 (A) DW is modified,
  (α) The number of DW is only one, then DW>EW
  (β) The number of DW are more than one, then CD
 (B) DW isn't modified ,
  (α) The number of DW is only one, then DW≡EW
  (β) The number of DW are more than one, then DW<EW

(II) DS is in TYPE II (DS includes FE),
 (A) DW is modified ,
  (α) The number of DW is only one,
   ρ_{FE} is '>' or '≡' ,then DW>EW otherwise CD.
  (β) The number of DW are more than one, then CD
 (B) DW isn't modified ,
  (α) The number of DW is only one, then DW ρ_{FE} EW
  (β) The number of DW are more than one,
   ρ_{FE} is '<' , then DW<EW otherwise CD.

CD denotes that DW-EW relation isn't extracted mechanically from DS. In this case, the extraction of DW-EW relation needs human support at this stage.

2.3 FEATURES OF FE

FE prescribes hierarchical relations(e.g. DW>EW, DW<EW, DW=EW, or DW≡EW) or whole-part relation(DW》EW). (e.g. On "【間脳(interbrain)】:…、脳(brain)の一部分 (a part of)。", the FE "の一部分" prescribes DW》EW explicitly.)

Besides these relations, another relation between DW and EW are prescribed by special FEs(e.g. "の下(under)"), which is called associative relation(R).

There are so many FEs that they are mainly divided into four patterns called functional patterns{FP: a brief notation}. FP is expressed by means of regular expression. FP is necessary for extracting FE and DW-EW relation information(i.e. information neccessary for deciding the DW-EW relations) assigned by the FE. FP also designates a place of DW in DS.

Main features of FP are as follows:
 (1) Type100 : 「…DW」・σ*・FW
 (2) Type200 : …DW・(の・FW)*
 (3) Type300 : …DW・P・FW
 (4) Type400 : …DW・など

Notes) σ* is arbitrary character string,
 (…)* is repitation of (…),
 P is special phrase(e.g. に対する),
 ・ is concatination symbol.

We got about one hundred seventy FWs. These are classified into two groups. In one group(contained 64 FWs), the FWs contain explicitly DW-EW relation information. In the other group(contained 105 FWs), some of the FWs contain usages of the EWs, which are also important to thesaurus.

We have constructed a FW dictionary which includes FP and DW-EW relation information corresponding to the FP.

3. SYSTEM FOR EXTRACTING DW-EW (HIERARCHICAL) RELATION

The system consists of the following four steps.

(1) Extraction of EW and DS

(a) Extraction of EW, its DS, the part of speech of the EW, the definition number of the DS from the dictionary.

(b) Transformation of the extracted DS to the ordinary Japanese sentence's form(called the normalized DS). Because several contents(meanings) are thrown into one DS by means of parentheses or dot '・' in the dictionary.

(2) Extraction of FE and DW-EW relation information

The FW Dictionary is used.

(a) When DS dosen't include FW, DS is in TYPE I.

(b) When DS includes FW and conforms FP, DS is in TYPE II.

(c) When DS includes FW but doesn't conform FP or when DS includes more than one FW, the DS is picked out as check data.Because it is difficult to distinguish between DW and FW or to extract DW-EW relation information mechanically.

(3) Extraction of DW and DW-EW relation information

A general word dictionary (containing about 75,000 noun words)[5]is used, in which the character strings of entry words were arranged in inverse order (from right to left). DWs are basically extracted by means of longest matching method, because there is ordinarily no space between two adjacent words in the Japanese sentence. In addition to this, there are the following problems.

(a) The 'hiragana' notation is often used(e.g. ものさし [物差し]).

(b) The names of animals and plants are described by 'katakana' (e.g. カエル [蛙]).

(c) The unknown(compound) words are often used.

(d) In some cases, the DS containes more than one DW.

The extructing procedure has to be constructed with regard to these ploblems.

The relation information are also extracted,that is, 'DW isn't modified' and 'The number of DW are more than one'.

When DW isn't extracted (that is,DW is neither 'katakana' string nor 'kanji' string nor any entry word in the word dictionary) from DS, the DS is picked out as check data.

(4) Decision of DW-EW relation

According to the conditions above-mentioned, DW-EW relations are decided.

When extracted relation information is ambiguous, DS is

picked out as check data.

4. EXPERIMENTAL RESULTS

A pilot system has been implemented on FACOM M-360(Nagasaki University Computer Center) and FACOM M-382(Kyushu University Computer Center) mostly by PL/1.

The experimental input data(2,824 DSs) in the first step, are the normalized DSs.

Table 1 shows the number of input, output and check data in each step and the number of correct and incorrect data in output data.

Table 2 shows the extracted DW-EW relations and the number of output data corresponding to the relations.

The experimental results are as follows:

(1) The ratio of TYPE I (2,374) to output data(2,711) is about 87.6%.

(2) The ratio of TYPE II (337) to output data(2,711) is about 12.4%.

(3) The ratio of output data(2,434) to input data(2,824) is about 86%.

(a) The ratio(called system precision) of correct output data(2,311) to output data(2,434) is about 95%.

(b) The ratio(called error ratio) of incorrect output data(123) to output data(2,434) is about 5%.

(4) The ratio of check data(390) to input data(2,824) is about 14%.

Table 1 The Number of Input Data, Output Data and Check Data in Each Step

	INPUT DATA	OUTPUT DATA (correct:incorrect)	CHECK DATA
(1) Extraction of FE	2,824	2,374(TYPE I) 337(TYPE II)	113
(2) Extraction of DW	2,711	2,502 (2,386: 116)	209
(3) Decision of Relation	2,386	2,318 (2,311: 7)	68
Result of Experiment	2,824	2,434 (2,311: 123)	390

Table 2 DW-EW Relations and the Number of Output Data corresponding to Each Relation

Relation	Correct		Dubious		Incorrect		Subtotal	
DW : EW	Type I	Type II	Type I	Type II	Type I	Type II	Type I	Type II
>	1963	71	3	0	0	0	1966	71
≡	120	103	0	0	0	0	120	103
<	24	11	0	1	0	0	24	12
≫	0	9	0	1	0	1	0	11
R	0	10	0	0	0	1	0	11
Subtotal	2107	204	3	2	0	2	2110	208
Total	2311		5		2		2318	

Most of incorrect output data occur in the step of extraction of DWs which are described by 'hiragana' notation, because of limitaions of the longest matching method.

The improvement of the results necessitates (a) analysis of the DSs, (b) reinforcement of the general word dictionary used for extracting the DWs.

5. CONCLUDING REMARKS

(1) The similer researches have been carried out about several English dictionarys(e.g.LONGMAN)[2][3], however there is scaresly any about Japanese dictionary.

(2) We have extracted automaticlly, DW<EW, DW≡EW, DW≫EW in addition to DW>EW as the DW-EW relations.

(3) Input data not suitable for conditions are picked out as check data in each step.

(4) There are a shortage of semantic information (e.g. lack of the adequate DW) in the dictionary because of assuming the human usage of the dictionary.

We have been investigating the followings.

I. Development of a system for utilizing the dictionary[7].

II. Development of a system for hierarchically structuring among entry words in the dictionary[6].

III. Development of a man-assisted system for constructing a practical sized semantic dictionary[4].

ACKNOWLEDGEMENT

We will like to thank the member of Yurumaru's laboratory in Nagasaki University, and in paticular, Mr. A.Uchida and Mr. K.Mizuno for their efforts of implementation.

REFERENCES

(1) S.Yokoyama: Preparation for the Data Management of a Japanese Dictionary, Bul. Electrotech Lab., Vol.41, No.11, PP.19-27 (1977.11)

(2) M.Nagao,J.Tujii,Y.Ueda,M.Taiyama: AN ATTEMPT TO COMPUTERIZED DICTIONARY DATA BASES, Proc. COLING80, PP.534-542 (1980.10)

(3) A.Michiels,J.Noel : APPROACH TO THESAURUS PRODUCTION, Proc. COLING82, PP.227-232 (1982.7)

(4) S.Yoshida,H.Tsurumaru,T.Hitaka : MAN- ASSISTED MACHINE CONSTRUCTION OF A SEMANTIC DICTIONARY FOR NATURAL LANGUEGE PROCESSING,Proc. COLING82, PP.419-424 (1982.7)

(5) K.Yoshimura,A.Yamasita,T.Hitaka,S.Yoshida : Automatic Extracting System of Technical Terms, NL Technical Report of IPS, 42-1 (1984.3)

(6) H.Tsurumaru,K.Mizuno,A.Uchida,T.Hitaka, S.Yoshida: Extraction of Hierarchical Relation between Words from Definition Sentence of Word, NL Technical Report of IPS, 45-5 (1984.9)

(7)H.Turumaru,A.Uchida: The Extraction and Organization of Information from the Ordinary Japanese Language Dictionary, Reports of the Faculty Engineering Nagasaki University, Vol.15, No.24 (1985.1)