Thesaurus Construction and Analysis Method for Dialogue Understanding
Subtitle: Thesaurus Construction for Dialogue Understanding

Hiroaki TSURUMARU, Shinichi HIGO and Takeo KIDO
Department of Electrical Engineering and Computer Science,
Nagasaki University, 1-14 Bunkyou-machi, Nagasaki 852, JAPAN
e-mail: turumaru@ec.nagasaki-u.ac.jp

1. Introduction

Many ellipses and demonstrative pronouns occur in dialogue. Generally speaking, the omitted words (or phrases) and the pronominal references are complemented by the use of common sense and discourse information. The definition of common sense is not clear but the thesaurus is regarded as an approximate model of the common sense. There are some thesauri such as " Bunr ui-Goi-Hyo(Word List by Semantic Principles) " and " Roget's Thesaurus ". H owever, these human-oriented thesauri are not always sufficient for natural language processing, because ambiguous relation between concepts exist (ref.[1]).

This study aims to clarify the method for constructing a thesaurus based on hierarchical relations such as upper/lower relation and part/whole relation, and intends to approach the problems of the application of the thesaurus to dialogue understanding.

Now, how and from what to obtain these hierarchical relations is one of the most important problems for constructing the thesaurus. Ordinary dictionaries include common sense on language. We have been studying how to acquire the word knowledge such as these hierarchical relations from the on-line Japanese dictionary, and then developing various programming systems for computer-aided thesaurus construction.

First, we mention that the thesaurus based on hierarchical relation is effective against the problem of ellipses and anaphora in dialogue. Second, explain on the thesaurus construction using a Japanese language dictionary. And then, briefly discuss on some of other results, e.g. a support system for improvements of the trial thesaurus, and the application of the thesaurus to the dialogue understanding.

2. Dialogue Understanding and Thesaurus

For complementing the omitted words and the pronominal references in dialogue sentence, the deep structure is often used. The case-frame and the thesaurus are necessary for extracting the deep structure from the sentence.

For example, in the following dialog sentence,

(1) フランス語話す？ (Dose ( ) speak French ?)

given one of the case-frames of ' 話す(speak) ' such as " agent is ' 人(human) ' and object is ' 言葉(language) ' " and obtained the upper/lower relation be tween ' フランス語(French) ' and ' 言葉(language) ' by means of thesaurus, o ne of the deep structures such as " agent is ' 人(human) ' and object is ' フランス語(French) ' " is inferred and then it is deduced from this deep str ucture that a case element of agent and a case relation of object (postpositional particle) are omitted. And furthermore, additional information such as discourse information are needed for complementing the omitted words. At this point, how to prepare the case frames of verbs is one of the most difficult problems. There are two standpoints concerning this problem. One is to prepare as fine case frame as possible. Another is to prepare examples of sentence level instead of sophisticated case frame. In either case the thesaurus which clarifies the hierarchical relation between words is important (ref.[2]).

Concerning the part/whole relation, for example,

(2) ぶどうを食べるとき，種は出しなさいね.

(When you eat a grape, spit out the seeds.)

Knowledge of a part/whole relation between ‘種(seed)’ and ‘ぶどう (grape)’ is necessary for inferring about what seeds, therefore the thesaurus is e ffective for obtaining such knowledge. Given the part/whole relation between ‘果物(fruit)’ and ‘種(seed)’ and the upper/lower relation betwee n ‘果物(fruit)’ and ‘ぶどう (grape)’ in the thesaurus, a part/whole relat ion between ‘種(seed)’ and ‘ぶどう (grape)’ is deduced by means of hierar chical relation with inheritance property .

3. Thesaurus Construction Using a Japanese Language Dictionary

We have been developing various computer-programming systems for the word knowledge acquisition from on-line Japanese language dictionary. In this section we pick up two main systems related with thesaurus construction. One of them is the system for extracting hierarchical relation between the entry word (EW) and the corresponding definition word (DW) from the definition sentence, i.e. the extracting system (ref.[4]). The core word expressing the central meaning in the definition sentence is called the definition word. Here we regard the upper/lower relation($>$), the part/whole relation($\ll$), the synonymous relation($\equiv$) and element/set relatio n($\in$) as hierarchical relation in a wide sense (ref.[6]).

Another is the system for thesaurus construction, i.e. the thesaurus system (ref.[3]). This thesaurus system is composed of two subsystem, the entry constructing system and the hierarchical structure searching system. By using the results of these works, a thesaurus has been made on an experimental basis, which has about 82,000 entry words (concepts) of noun. Here we regard a sense given by a definition sentence as a concept. Input data for the entry constructing system are the EW-DW relation data obtained from definition sentences of noun entry words(about 55,000 words) in Shin-meikai-kokugo-jiten(Sanseido) by the use of the above-mentioned extracting system.

Fig.1 and Fig.2 show two examples of a entry : ‘金 (gold) ［きん］ (1110) ’ and a hierarchical structure : ‘砂金 (gold dust) [さきん](0010) ’ in our trial thesaurus. The entry word (concept) is composed of three element, Kanji-notation, reading(Kana) and meaning (definition) number in MRD. For convenience' sake English equivalents are given in parentheses after Kanji-notation. Here, the average depth of the conceptual hierarchy is 3.2.

4. Some of Other Results and Discussion

(1) Concerning improvements of the thesaurus, we have been developing various programming-tools for thesaurus construction support. One of them, we call it xthtree, is shown in Fig.3, which indicates the hierarchical structure of given word(concept) on GUI, and is installed on Sun Sparc station. This tool is very useful, for example, for assigning the superordinates to the local maximal words having no connection with the superordinates, and so on.

(2) The part-whole relation is useful in analyzing the dialogue sentences. But only hundreds of entry words have part-whole relation in our pilot thesaurus. Then we have proposed three types of logical definitions of part-whole relation, and a logical method for

金 (gold) ［きん］ (1110)
＜： 金属 (metal) ［きんぞく］ (0010)
金属元素 (metallic element) ［きんぞくげんそ］
： 元素 (element) ［げんそ］ (0010)
＞： 砂金 (gold dust) ［さきん］ (0010)
： 純金 (pure gold) ［じゅんきん］ (0010)
： 焼き金 (pure gold) ［やきがね］ (0310)
≡： ゴールド (gold) ［ゴールド］ (0010)
： 黄金 (gold) ［こがね］ (0010)

Fig. 1 An Example of Our Trial Thesaurus entry

砂金 (gold dust) [さきん] (0010)
＜：金 (gold) [きん] (1110)
＜：金属 (metal) [きんぞく] (0010)
＜：鉱物 (mineral) [こうぶつ] (0010)
＜：無機物 (inorganic substance) [むきぶつ] (0010)
＜：物質 (matter) [ぶっしつ] (0110)
＜：もの (thing) [もの]
＜ 金属元素 (metallic element) [きんぞくげんそ]
＜：元素 (element) [げんそ] (0010)
＜：もの (thing) [もの]

Fig. 2 An Example of the Hierarchical Structure

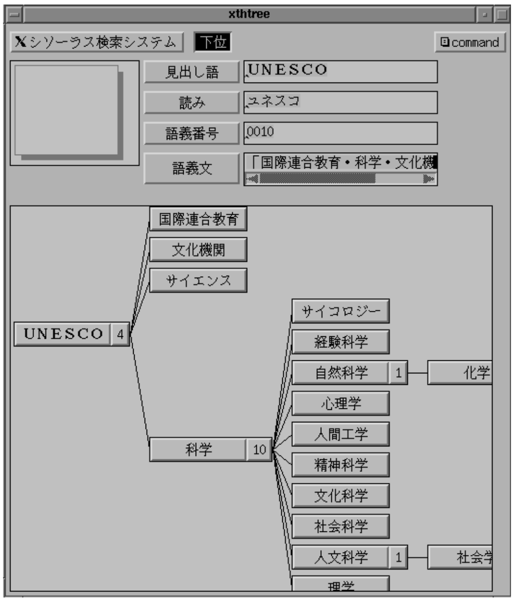

Fig.3: An Example of xthtree Output

deducing the new words-pairs of part-whole relation from the given words-pairs of part-whole and others relation.

(3) Concerning the application of the thesaurus, we have also been studying on an algorithm for the inference of the omitted words or phrases in dialogue using the thesaurus and IPAL basic verbs dictionary in order to verify the validity of the trial thesaurus. The outline of the algorithm is as follows;

[a] morphological analysis of input sentence,

[b] extraction of the deep structure (case structure),

[c] recognition of the existence of the omitted words or phrases,

[d] inference of the omitted words or phrases.

Similar algorithm may be used to handle pronominal and anaphoric reference.

(4) Toward the word knowledge base construction, we have been studying on the word knowledge acquisition from on-line Japanese language dictionary. The word knowledge base is a kind of the thesaurus in wide sense, which includes semantic relational information among words (concepts) from various points of view. The definition words (concepts) in definition sentences are qualified and specialized from various points of view. We discuss on how to extract such semantic information involved in the definition sentence (ref.[5]).

References

[1] H.Tanaka, K.Nishina : Construction of a thesaurus based on Superordinate/ Subordinate Concept[][], IPSJ SIG Notes, NL64-4, Vol.87, No.84, pp.25-4 4 (1987)

[2] S.Kurohashi, M.Nagao : On the Effectiveness of Example-based Approach to Semantic-marker Approach in the Selection of Japanese Case Frames, IPSJ SIG Notes, NL91-11, pp.79-86 (1992)

[3] H.Tsurumaru, et al. : A Study on Thesaurus Construction by Using Japanese Language Dictionary,IEICE Technical Report, NLC93-58, Vol.93, No.367, pp.29-36 (1993)

[4] H.Tsurumaru, S.Yoshida : An Approach to Extraction of Semantic Relations from an Ordinary Japanese Language Dictionary, 47th FID(International Federa tion for Information and Documentation) Conference and Congress, pp.138-142 (1994)

[5] H.Tsurumaru, et al. : A Study on Thesaurus and Sematical Classification o f Words Based on Japanese Language Dictionaries, IEICE Technical Report, NLC95-44, pp.31-38 (1995)

[6] H.Tsurumaru, et al. : Functional Expression of Definition Sentences in Ja panese Language Dictionaries, IEICE Technical Report, NLC95-65, Vol.95, No.429, pp.47-54 (1995)

keywords: thesaurus, semantic dictionary, word knowledge base, conceptual hierarchy, dialogue understanding, natural language processing