Acquisition of Kowledge Data by Analyzing Natural Language

Yasuhito Tanaka

Himeji College
1-1-12 Shinzaike Honmachi
Himeji City Hyogoken
670 JAPAN

Sho Yoshida

Kyushu University
6-10-1 Hakozaki Higashiku
Fukuoka City Fukuokaken
812 JAPAN

1. Introduction

Automatic identification of homonyms in kana-to-kanji conversions systems and of multivocal words in machine translation systems cannot be sufficiently implemented by the mere combination of grammer and word dictionaries. This calls for a new concept of knowledge data. What the new knowledge data is and how it can be acquired are mentioned in the paper. In natural language research, active discussion has been make within the framework of knowledge and samples of knowledge.

2. Phases of Natural Language Study and Knowledge Data

〔phase 1〕

In the initial phase when natural language had not been fully clarified, the authors made an attempt to analyze the language with the help of word dictionaries and grammar alone, and to build a new system from the results. We, however, encountered a great number of homonyms in kana-to-kanji conversion, as well as multivocal words and other problems in machine translation. This was because the grammar and dictionaries were too simple. In other words,

$$\omega_i \rightarrow f_T(\omega_i)$$

where ω_i is a word and f_T is a conversion system.

〔phase 2〕

There are some possible means for solving the problems encountered in Phase 1. They include the following: analyze the word frequency, limit the number of words in use, limit the contents of the words in use, limit the fields and limit sentences. This helps remove, or at least minimize, homonyms, multivocal words and ambiguities. A system with limitations, however, involves too many difficulties to be readily accepted.

〔phase 3〕

In vocabulary, no words are used independently. One should define and use words by clarifying the characteristics, coverage and conditions of words. This may be expressed as follows:

$$\omega_i \mid P_{i1}, P_{i2}, P_{i3}, \cdots\cdots, P_{in}$$
$$\rightarrow f_T(\omega_i \mid P_{i1}, P_{i2}, \cdots\cdots, P_{in}) = yj$$

where ω_i is a word; and P_{i1}, P_{i2}, $\cdots\cdots$, P_{in} are the limitations of ω_i. The set of P_{i1}, P_{i2}, $\cdots\cdots$, P_{in} should be low in number and simple.

Different approaches have already been proposed as to what is necessary for clarifying the conditions for the coverage of words in Phase 3. Some examples are semantic markers, semantic categories and thesaurus. However, the authors put emphasis on the relationship of words, though recognizing such other forms of approach.

〔Knowledge Obtained from Words〕

Words and sentences provide knowledge shown in Table 1, from simple to complicated in order. How much of them are usable as dictionaries (books)? How much are usable as machine-readable dictionaries?

In Table 1, machine-readable dictionaries have already been established for 1. "Word-related Attributes". Classified vocabulary lists have been prepared for Thesaurus Structure in 2, but there is almost nothing for the other attributes.

Therefore, an attempt to build a sophisticated system would be unsuccessful if no basic knowledge data is available.

Table 1. Knowledge Obtained from Words

1	Word-Related Attributes
1.1	Word Attributes Words, parts of speech, pronunciation, accent and kana representation
1.2	Long Unit Words and Technical Terms
2	Word-to-Word Attributes (1) (Preconditions)
2.1	Broader/Narrower Rank Relationship (Thesaurus Structure)
2.2	Antonyms and Negatives
2.3	Partial/Whole Relationship
2.4	Sequential Relationship
2.5	Comparative Relationship (size, height)
3	Word-to-Word Attributes (2) (Preconditions)
3.1	Case Relationship
3.2	Relationship Based on the Synchronism of Sentence Components
3.3	Idiomatic Expressions
4	Word-to-Word Attributes (3) (Preconditions)
4.1	Association-Based Relationship
5	Sentence Relationship
5.1	Sentence-to-Word Relationship
5.2	Sentence-to-Sentence Linkage

3. Is it Possible to Define Word Coverage and Conditions?

Since every word may be linked with an infinite number of words, one may wonder if it is impossible to define word coverage and conditions. It would also take a great deal of time and trouble to examine all the words that are almost infinite in number. If, however, one actually examines some of the words, we will fine that every word is linked with a limited number of words.

Table 2. Each Word is Linked with a Limited Number of Words

001	Denwa o kakeru	018	Denwa o migaku
002	Denwa o kiru	019	Denwa o ukeru
003	Denwa o mochiageru	020	Denwa o tochosuru
004	Denwa o kowasu	021	Denwa o kakenaosu
005	Denwa o nigiru	022	Denwa o motsu
006	Denwa o motsu	023	Denwa o motaseru
007	Denwa o kairyosuru	024	Denwa o kiku
008	Denwa o tsukuru	025	Denwa ga naru
009	Denwa o seisakusuru	026	Denwa o tsutaeru
010	Denwa o kumitateru	027	Denwa de hanasu
011	Denwa o kaisetsusuru	028	Denwa de renrakusuru
012	Denwa o hiku	029	Denwa ni deru
013	Denwa o tekkyosuru	030	Denwa no koe
014	Denwa o uru	031	Denwa no buhin
015	Denwa o hanbaisuru	032	Denwa no kane
016	Denwa o kau	033	Denwa no ryokin
017	Denwa o konyusuru	034	Denwa no beru
		...	.....

Take the word, "denwa" (meaning telephone in English) as example. It has a limited number of characteristics such as a means of communication, a substance, a place, and so on. The function as a means of communication and the characteristics of the telephone are unique to the telephone.

In this particular meaning, it is a simple and finite task to count the different relationships between words. It is, however, extremely difficult to count the word-to-word relationships for the general meanings such as a substance and a place. However, it is possible to organize the major relationships without much effort.

The only approach would be to tabulate the words having peculiar relationship to specific words and the relationship between specific words and words of high usage frequency, and thus use system-preset defaults for other words.

Such words as "takai" (meaning high) and "utsukushii" (meaning beautiful) are used frequently. In some cases, therefore, it is difficult to determine the coverage and conditions of individual words. Such cases must include general grammar and the meaning of words most frequently used in "takai" and "utsukushii" must be defined in the word-to-word relationships together with the conditions of any special meanings of high and beautiful. Words of low frequency and high frequency are dependent on individual rules (conditions based on word-to-word relationship) and general grammar.

4. Aquisition of Knowledge data

4-1 Method by extracting and segmenting Kanji strings

Authors and S. Mizutani of Tokyo Women's Christian College mechanically extracted the four-character kanji strings from some JICST abstract files, extracted meaningful concept combinations from them, and classified them into 45 categories. Thus, after examing 78,000 four character kanji strings, 28,000 different kinds of knowledge data was obtained from 32,000 types of kanji strings in net. This approach is promising because it allows easy expansion in quantity and consists mainly of mechanical processing. In addition, an attempt to reorganize a total of 887,000 data, 200,000 in net, is under way. An attempt to apply the same idea to the three and five character kanji strings is also being made.

〔Example〕

Tが／をPv　会長・辞任　　Tにより Pv　　写真・判定

T ga/o Pv kaicho-jinin T ni yori Pv shashin-hantei

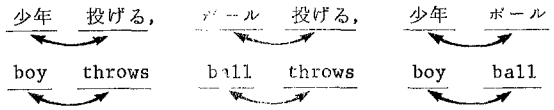
4-2 Aquisition of knowledge data based on word-to-word synchronism

In this case, Japanese sentences should be analyzed to extract connections and obrain knowledge data. Analysis, however, takes a great deal of time and involves ambiguity. This suggests an approach which simply extracts independent words handles them, if present in the same sentences, as candidates for knowledge data because of mutual relation, and extracts meaningful knowledge data based on frequency and manual procedure.

〔Example〕

少年がボールを投げる

A boy throws a ball.



To extract such relationships and reorganize them sequentially.

4-3 Aquisition of knowledge by syntax analysis

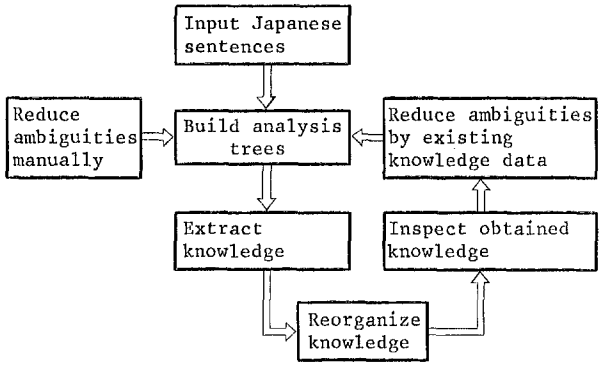
This is a method which analyzes the sentences that are input with word dictionaries and grammar, builds syntax elements, extracts connections from them, and thus, extracts knowledge data. This method is theoretically pertinent but not practical because it leaves the problem of a tremendous increasing number of syntax trees. It would be very effective if a great amount of knowledge data were available and some of the ambiguities from syntax analysis were reduced. Since, in Japanese sentences, long-unit terms are used as they are, it is also necessary to sub-categorize them automatically. Otherwise, a variety of knowledge data would result and it would be troublesome to systematize knowledge data.

XX daigaku (university) , XY daigaku ⇔ daigaku

general terms　　basic concept term

This sort of concept was presented by some manufacturers engaged in the research and development of machine translation. It is however only in the planning stage. There has been no news reporting that knowledge data is effectively available.

Fig. 2. Method of Obtaining Knowledge by Syntax Analysis



5. Applications of Knowledge Data

(1) To develop high quality Japanese word processors.

(2) To improve the quality of machine translation.

(3) To reduce the ambiguities of syntax analysis.

(4) To apply knowledge data in handwritten character and voice recognition.

6. Conclusion

The systemization, aquisition and construction of knowledge data are a step towards the next jump in Japanese processing systems. Indeed, the knowledge data still has a number of problems to be solved, but prospects for the future are rather bright.

References

(1) Inanaga & Konishi, Terms for Computer-Based Processing of Kana Characters AL 76-39 (in Japanese). Material of Engineering Workshop of Electronic Communications Society, 1976

(2) Shizuo Mizutani, Overview of Word Structure (in Japanese). Iwanami Japanese Language Dictionary (3rd Edition), March 1980

(3) Masaaki Yamanashi, Meaning and Knowledge Structure: Theoretical Study of Meaning Expression Models from Linguistics (in Japanese). Mathematical Science No. 240, June 1983.