Maintenance of Machine-Readable Dictionary

Yasuhito TANAKA
Hyogo University

2301 Shinzaike Hiraoka, Kakogawa, Hyogo
675-01 JAPAN
TEL : +81－794－27－5111 FAX : +81－794－27－5112

Abstract

This article discusses various problems related to machine-readable dictionaries. The author focused our attention on the corpus as a means of finding entries. Furthermore, we discusses these problems from the standpoint of users and also from the standpoint of system development. At the same time, the questions of verification of conversion accuracy and overall evaluation of developed dictionaries were studied.

0) Introduction

Machine-readable dictionaries have been compiled by companies, institutes, software houses, and others. No machine-readable dictionary maintains its usefulness without revision. Unless it is continuously renewed or maintained, it quickly becomes outdated. Here we will cite various problems related to the maintenance of dictionaries and introduce the results of tests regarding the solution of the above-mentioned problems.

1) Problems related to maintenance of dictionaries

i) Dictionaries become dated year by year.

Although a machine-readable dictionary may be up-to-date when it is completed, it becomes progressively more outdated year by year. For instance, one can easily recognize that sentences written 10, 20 or 50 years ago are old in content, concepts, diction, etc. Similarly, dictionaries get old.

ii) Assurance and maintenance of personnel for renewal

Personnel must be secured for the maintenance of a machine-readable dictionary. Such personnel must be experienced in the processing of natural language and have an adequate knowledge of linguistics. Furthermore, they must continuously be able to complete a fixed amount of work within a given period. This is a very difficult task.

iii) How to obtain data for maintenance

If a product using a machine-readable dictionary is offered on the market, it is possible to obtain data for renewal just by gathering complaints from users.

Another way is to extract words from the corpus, and refer such words to a machine-readable dictionary. If there are words that are not found in the dictionary, new words that appear with high frequency may be added to the dictionary. This positive method is important.

2) How can data for maintenance of a machine-readable dictionary be obtained, and how should unknown words be handled?

In developing a system for processing natural language, it is necessary to perform the following three types of maintenance for a machine-readable dictionary.

(1) Maintenance for solving problems

If a new word (unknown word) is found, its meaning should be determined based on its outward characteristics. It is impossible to deal with new words in the course of sentence processing.

(2) Maintenance according to users' demands

New words are added to a machine-readable dictionary according to users' demands.

It is possible to have users make proposals. For instance, an electronic bulletin board can be provided in a communication network system to collect data. The degree of success of this method depends greatly on whether the information receive rewards or some kind of privilege.

If the consent of users can be obtained, words, compound nouns, idiomatic expressions, technical terms, coincidental relations, etc., that the users use and add to their machine-readable dictionaries, may be collected so that words, etc., of common use can be selected and incorporated in a dictionary for a natural language processing system.

(3) Preventive maintenance

There is also a method of continuously extracting new words from a large corpus and registering them in a machine-readable dictionary. By this method, it is possible to analyze a number of words according to the frequency with which they appear. It is also possible to use corpuses devoted to different areas. This is a positive and important task in the maintenance of machine-readable dictionaries.

It is necessary to find and eliminate as many detects as possible before a product is put on the market. The correction of defects after shipment of the product requires tremendous costs.

Here we will focus on (3) Preventive maintenance.

3) Extraction of words from a corpus

3-1)

In extracting words from a corpus, it is possible to analyze sentences and extract words. However, this would require a great deal of time and effort for processing, and a completed machine-readable dictionary would be needed. Unfortunately, a dictionary will never be complete.

Another way is to mechanically separate text into words and then to select words. Words can be extracted incorrectly when this method is used, so care should be taken in incorporating them into a dictionary. It is possible to refer words extracted from a corpus to those already registered in a dictionary and to consider adding unmatched words as candidate words to be newly incorporated in a dictionary.

This method is advantageous in that it depends to a considerable extent on mechanical processing, and a large number of candidate words can be found in this way.

Thus, we extracted three-Chinese-character words from a data file containing the full text of the Asahi Shimbun over a period of a year.

Code	№ of different words	Total number of words	Code
10	35,623	406,827	10…common nouns
80	10,427	96,946	80…proper nouns (names of persons)
90	3,312	24,536	90…proper nouns (place names)
95	763	3,292	95…numerals
99	17,941	112,711	99…others
Total	78,066	644,312	

3-2) Checking of words (examples)

One way to check words is to check words extracted by a mechanical method. Here we will introduce another method. A large number of four-Chinese-character words are divisible into combinations of two two-character conceptual words. These two-character words should be included in a dictionary. We examined whether such two-character words are included in a machine-readable dictionary.

Machine-readable dictionary

	№ of different 2-character words	Total number of different of 2-character words
Yes	10,512	794,649
No	3,336	17,740
Total	13,848	812,389

The above table shows that since two-character words are basic words, the machine-readable dictionary covered 97.8 percent of the total number of different two-character words.

We conducted a similar analysis of the data provided by the Japan Scientific and Technical Information Center, and obtained the following results.

Machine-readable dictionary

	No. of different 2-character words	Total number of different of 2-character words
Yes	8,3711	519,820
No	9,3821	11,889
Total	17,753	1,631,709

It is clear that there are more unmatched words for the data obtained from the Japan Scientific and Technical Information Center.

This means that the evaluation of machine-readable dictionaries differs depending on the main areas for which they are developed.

Five-character words were extracted and divided into two-character and three-character words. Of these, three-character words were checked with a machine-readable dictionary.

Machine-readable dictionary (3-character words contained within strings of five characters)

	No. of different 3-character words	Total number of of 3-character words
Yes	11,034	109,051
No	2,492	8,311
Total	13,526	117,362

The above figure shows that the machine-readable dictionary covered 92.9 percent of the total data.

Similarly, two-character words contained within strings of five characters found in the Asahi Shimbun file were checked with a machine-readable dictionary.

A very good result was obtained, as the machine-readable dictionary covered 96.1 percent of the total data.

Machine-readable dictionary (2-character words contained within strings of five characters)

	No. of different 2-character words	Total number of of 2-character words
Yes	5,573	113,947
No	843	3,415
Total	6,416	117,362

Machine-readable dictionaries are maintained by analyzing data obtained in this way.

4) Four-character and five-character words with furigana

An attempt was made to determine how many of the four-character words that can be divided into two two-character words can have furigana (phonetic transcriptions in kana written at the side). The following table shows the number of four-character words which can have furigana entirely.

Four-character words with furigana

	No. of different four-character words	Total number of four-character words
Yes	68,465	377,062
No	11,212	29,133
Total	79,677	406,195

A similar attempt was made for five-character words that can be divided into two-character and three-character words, or vice versa, and the number that can have furigana was determined.

Five-character words (2•3)

	No. of different five-character words with furigana	Total number of five-character words with furigana
Yes	14,385	49,390
No	3,320	7,881
Total	17,705	57,271

Five-character words with furigana (3•2)

	No. of different five-character words with furigana	Total number of five-character words with furigana
Yes	18,181	52,043
No	3,895	8,050
Total	22,076	60,093

When providing furigana to Chinese characters it should be remembered that the first consonants in the latter parts of four-character and five-character words tend to be voiced by liaison after the first parts. Ex. Kabushiki Kaisha (Gaisha) Data obtained in this way can be excellent material for addition to machine-readable dictionaries.

3-3) Checking with a new file

Three-character words were extracted from the file containing one year of the full text of the Yomiuri Shimbun and were added to those extracted from the Asahi Shimbun file.

Three-character words from the one-year Yomiuri Shimbun file

Code	No. of different 3-character words	Total No. of 3-character words
10	19,960	432,426
80	3,571	91,261
90	1,837	31,221
95	443	6,948
99	6,969	57,219
None	66,121	234,208
Total	98,901	853,283

This table shows that about 66,000 different three-character words out of the total of about 99,000 are codeless. This means that about 67 percent of the total different words and 27.4 percent of the total number of words are unmatched data. This increase in unmatched data is probably due to the mechanical method of extraction and also to the fact that a large number of the words contain numerals. It is also a fact, however, that there are more words that are required to be registered. So far as common nouns (Code 10) are concerned, about 20,000 words are matching words, and the total number of such words is about 432,000. Furthermore, 66,000 different words are code-less, and they appear a total of about 234,000 times, or an average of 3.5 times per each words. It will be necessary in the future to analyze these words one by one and register them in the dictionary. About 66,000 different word s are now being analyzed.

3-4) Method of analysis

Checking with existing files is the first step in analysis. In this analysis, only about 30 percent are matching words. Therefore, the following empirical method should better be followed.

However, it should be remembered that there are exceptions to any rule.

(1) Three-character words containing one of the following characters are place names (90).

県、市、区、郡、町、村、字、省 and three-character words of which the last character is following 浦、駅、沖、河、海、街、岳、橋、郷、局、圏、原、湖、江、港、崎、山、坂、寺、州、沼、川、線、台、沢、谷、地、島、峠、灘、岬、野、路、湾

(2) Three-character words containing one of the following characters are numerical expressions (95).

一、二、三、四、五、六、七、八、九、十、拾、百、千、万、億、兆

(3) Three-character words containing one of the following characters as the last character are names of persons or proper nouns (80).

～氏、～子

(4) Words beginning with the following code are "other nouns" (99).

案外、以下、以外、以上、以後、以前、以来、依然、一応、一括、一見、一言、一向、一時、一瞬、一生、一切、一層、一体、一度、一回、一般、一晩、一番、一部、一歩、一面、一目、一律、何人、何度、過去、回、極力、月末、現在、現代、今後、今回、今後、今月、今春、今週、今度、今日、今年、再三、再度、最近、最終、最大、最低、最高、昨年、氏

(Characters at the beginning of words)

時折、時代、若干、最終、終日、週間、週末、十分、従来、順次、将来、場合、常時、随分、数日、数年、世紀、世代、絶対、先月、先週、先日、戦後、戦前、前回、前後、前日、前年、前夜、全国、全身、全然、全部、全面、早速、相当、即刻、多少、多数、対応、断然、直接、通常、程度、当時、当初、当然、当分、当日、当面、到底、同年、突然

There are also other words with various characteristics, and we are now analyzing them. When the two files are integrated, we obtain the following table.

Three-character words in the files of Asahi Shimbun and Yomiuri Shimbun Files

Code	No. of different 3-character words	Total No. of 3-character words
10	35,623	839,253
80	10,427	188,207
90	3,312	55,757
95	763	10,240
99	27,940	169,930
None	66,121	234,208
Total	144,186	1,497,595

"None" denotes that code numbers are now being added.

When the analysis of 66,000 different words has been completed, a better file of analysis will be obtained. Furthermore, 60-70 percent of the different three-character words can be automatically coded.

We have discussed various problems related to dictionary maintenance from the standpoint of compilers and managers of machine-readable dictionaries. However, it is also necessary to view these problems from the standpoint of users.

4) Satisfaction of users

Experiences of people engaged directly or indirectly in the development of machine-readable dictionaries show that the demands of their users are satisfied in different degrees.

(1) When no machine-readable dictionary was available and different organizations had to compile their dictionaries, the existence of a machine-readable dictionary itself was highly ⁄ appreciated. Even a small-scale dictionary was welcomed.

A dictionary with 500,000-100,000 entries was appraised by users.

However, as the number of users increased and as the utilization areas of machine-readable dictionaries expanded, users became more demanding.

(2) They wanted an increase in the number of entries. Furthermore, they wanted to have conditions for their usage clarified so that they could select suitable words more easily. Thus, they wanted a dictionary with 200,000-300,000 entries.

We are now at this stage of development of machine-readable dictionaries.

(3) In addition to the demand for a largerscale dictionary, there will also be a demand for more detailed descriptions of each entry, semantic contents, origins and related information on words, as well as for clarification of usage of words. Machine-readable dictionary will come to be questioned as to their scale and c ontents, and a dictionary with about 1 million entries will be desired.

5) Viewpoint of system development

(1) Before there was a machine-readable dictionary, or when there was one of only a limited scale, system development was processed only by rules, and exceptions were also dealt with by rules. It soon became clear, however, that there were many problems in this method, and systems development got nowhere.

(2) Expansion of scale of dictionary and grouping of entries

The method of expanding a machine-readable dictionary to a certain extent for word-processing proved to be more efficient, and it became clear that there was no need to deal with exceptional entries by rules one by one.

However, there was concern about the extent to which dictionaries could be expanded. Against this background, attempts to group similar entries began.

Active studies were made on markers and their expansion, and also on a thesaurus.

(3) Development of a large-scale dictionary

In order to develop a large-scale dictionary, it is necessary to collect entries for it from a corpus. It has become possible to obtain a large corpus in the form of newspaper files thanks to the electronic photocomposing systems which have been extensively adopted by newspaper publishers. Furthermore, systems for electronically processing manuscripts have been developed for book and magazine publishing houses, and as a result, electronic books and magazines have made their debuts. However, these developments also copyright problems.

Furthermore, the method of dictionary development is changing from a method based on personal work to machine compilation, or a combination of machine compilation and review by personnel. Development of a large-scale dictionary depends to a large extent on computers, and for this purpose, methods of data extraction by statistical processing, the n gram system, and other methods have been developed. However, since the extraction of entries and contents for a dictionary by these methods is mechanical, it should ultimately depend to a great extent on human judgment.

It has become possible to compile a large-scale dictionary thanks to the following three factors.

(1) Research and development efforts have been made to resolve ambiguities, and medium-scale machine-readable dictionaries have been developed. These developments have helped to automate the compilation of machine-readable dictionaries.

(2) Large-scale corpuses have been made available.

(3) Various statistical methods have been developed.

6) Verification of contents of dictionaries

While it is important to compile a large-scale dictionary, it is also necessary to take note of the fact that errors are accumulated. In this context, it is important to consider how to verify the content of each entry in the dictionary.

(1) Verification by personnel

This requires a many people, and the ability of verifiers should be examined.

(2) Method of mechanical inspection

Various dictionaries are compared partially, do to find errors.

(3) To evolve an experimental system for verification

It takes time to evolve such a system, and a lot of time is needed before a dictionary is actually tested.

(4) A large-scale dictionary is partially completed, and entries in the partially completed part are classified and verified.

(5) Others

It is necessary to study the systems to inspect machine-readable dictionaries.

7) Overall evaluation of dictionaries

According to the author's experience in a project carried out around 1972 to convert kana characters to Chinese characters for the names of persons, the development costs required for conversions with accuracy rates of 80, 90, 95 and 99 percent rose from are to two, three and for times, respectively. Although these differences in conversion rates were not so large relative to the costs, they were of great significance. In connection with this, we made the following conversion calculations. The overall evaluation of 90 and 95 percent appears to be a good evaluation result. However, by multiplying the overall evaluation rates by large negative numbers, we can change evaluation points substantially.

(1) $70+(30\times(-5))=-80$

(2) $80+(20\times(1-5))=-20$

(3) $90+(10\times(-5))=\ 40$

(4) $95+(5\times(-5))=\ 70$

(5) $99+(1\times(-5))=\ 94$

The above shows that tremendous costs (amount of work) are needed to correct wrong conversion results. Therefore, the above evaluation equation appears to assume reality.

In actual fact, development costs doubled when the conversion accuracy rate was raised from 90 to 99 percent. This is what this writer actually experienced around 1972 in developing a kana-Chinese character conversion system for names of persons.

The evaluation of a machine-readable dictionary differs according to whether a conversion accuracy rate is based on the examination of discrete words or of compound words, and also according to whether a discrete system (only form element analysis) or a system as a whole is evaluated.

An adequate evaluation method of a natural language processing system has yet to be evolved.

We should be strict in evaluating the present state of affairs, but optimistic about the future.

Furthermore, it will be necessary not only to turn our attention to the number of entries and coverage, but also to continue our efforts to upgrade the quality of entries and their contents.

8) Future tasks

In the above we have made a proposal concerning a policy on data to be added to dictionaries.

Together with an increase in the number of entries in a dictionary, the improvement of the quality of entries (expansion of contents) is desired. It is also desired to clarify the usage of words and word associations, and further expand contents concerning upper-ranking and lower-ranking words, antonyms, associated words, technical terms, idiomatic expressions, etc.

9) References

Yasuhito TANAKA : Maintenance of Machine-Readable Dictionary, 48th (for the first half of 1994) National Conference of the Information Processing Society, 3Q-1, March 23, 1994

Yasuhito TANAKA : Maintenance of Machine-Readable Dictionary (Part 2), 49th (for the first half of 1995) National Conference of the Information Processing Society, March 15, 1995·