Language, Information and Computation (PACLIC12), 18-20 Feb, 1998, 310-315

Machine-Readable Dictionary Headwords

Yasuhito Tanaka † Kenji Kita ‡
† Hyogo University ‡ Tokushima University

Selecting headwords is an important problem in developing a dictionary. We have analyzed several machine-readable dictionaries and large volumes of corpora in our aim of developing an extensive machine-readable dictionary with one million headwords. This paper shows analysis and comparison of results and discusses how we collect, sort and allocate information to the dictionary headwords.

1 Introduction

One vital theme in natural language processing is how to produce a good machine-readable dictionary. For this, how we collect, sort and allocate information to the dictionary headwords is critical, and this is what I shall be discussing in this paper.

2 Collecting Machine-Readable Dictionary Headwords

The following are fundamental requirements in collecting machine-readable dictionary headwords.

(1) We should analyze the corpora and adopt headwords that appear frequently. Crucial to this is a clear standard on the frequency point at which a headword is adopted and below which it is rejected.

(2) We have to decide whether to use long or short terms as headwords. Generally in natural language processing, long terms are used to reduced ambiguity. The problem with this, though, is that using long terms will make the dictionary much more voluminous because of the greater number of words needed. Considering the importance of clarity, however, this cannot be helped.

(3) Certain words form natural groups, and all words belonging to these groups should be included; for example, prefectures, days of the week, months of the year, and planets in the solar system. Including just one of these words in their respective groups without the others will create functional problems.

(4) Collation of synonym and antonym dictionaries. Synonym and antonym dictionaries have to be checked against each other to ensure there are no inconsistencies between the two, such as an entry in one but no corresponding entry in the other. Where there is no corresponding entry, the reason must be clearly shown.

(5) The idea behind dictionaries for machine processing and that behind dictionaries that people can thumb through are different, so we have to be aware that different terms are used. It is often best to include in machine dictionaries even terms that are easy for people to understand and to break down into their key elements.

For example, :r6HCf(on duty, at work) and F17rKv(the end of the same month).

These terms are not used as headwords in dictionaries for use by people.

(6) Collation with other machine-readable dictionaries. Machine-readable dictionaries sourced from normal dictionaries that people use to read and understand are not suitable for machine processing. This is because people are able to understand and infer as necessary, whereas this is not possible in machine processing.

†† Hyogo University, 2301 Shinzaike Hiraoka Kakogawa, Hyougo, Japan 675-01.
E-mail: yasuhito@humans-kc.hyogo-dai.ac.jp
‡ Faculty of Engineering, Tokushima University, Tokushima, Japan 770. E-mail: kita@is.tokushima-u.ac.jp

(7) Selection of headwords. Writings by people who have developed normal dictionaries indicate that of all the possible headwords collected for the dictionary, only about a half or even fewer were actually used. We, too, have to think along these lines for machine-readable dictionaries.

Headwords may become obsolete or fall out of common use as the language evolves over time, so we must always analyze the frequency with which they are used. We must constantly monitor any changes in usage and frequency. Non-standard usage, young people's language, lyrics and proverbs are all a reflection of their age and change from generation to generation, or even more quickly. The following data bear this out.

=qL>	4)G/	?75,	:<=I	<)O78l?t
L@2r9q$1	1943	8,000	-	73,000
L@2r9q$1(2-Dj)	1952	7,000	14,000	66,000
;O>JF29q$1	1960	5,000	8,000	59,000
N_7W		20,000	22,000	-

Source: Takenori Kenbo, Jisho wo Tsukuru, p19, Tamagawa Sensho

(8) Variations in listings. In normal dictionaries variations in the listings can be standardized into a single entry, but for machine-readable dictionaries, in most cases processing is simpler if the number of headwords is increased.

This does, however, have its drawbacks in that the dictionary will be larger, and there will be some duplication of items when printed.

3 Collation with Data-Based Machine-Readable Dictionaries

Next I looked at data-based machine-readable dictionaries using dictionaries produced by two private companies.

3.1 Collation of three-character compounds

For this collation I used five years' worth of data from the Nihon Keizai Shimbun, namely three-character kanji compounds extracted using changes in character kind.

Some of the three-character compounds in the data were extracted incorrectly or by mistake, so the totals do not correspond.

Company A was set up with finance from several major manufacturing companies and the government, and has produced several machine-readable dictionaries. For this collation, I used the company's Japanese language dictionary. It contains about 200,000 words.

Company B is a well-known for its word processing software, and for the collation I used the seventh edition of its dictionary, containing about 130,000 words. The dictionary has since been further revised and is now at its tenth edition. The collation revealed the following.

Tests using different data		%
Total number of distinct words	318,472	100%
Company A dictionary		
Matching words	11,157	3.50%
Temporal adverbs	8	
Adverbs	9	
Common nouns	10,272	
Proper nouns	211	
Company B dictionary		
Matching words	35,517	11.15%
Common nouns	27,894	
Noun adjectival verbs	1,253	
Noun anomalous conjunction of the "sa" series	1,572	
Proper nouns (places)	4,021	
Proper nouns (persons)	369	
Proper nouns - general	181	
Proper nouns (organizations)	218	
Numerals	3	
Adverbs	6	

Results using total data		
Frequency of word use	8,382,228	100%
Company A dictionary		
Matching words	2,891,570	34.50%
Temporal adverbs	21,398	
Common nouns	2,775,161	
Adverbs	1,404	
Proper nouns	93,607	
Company B dictionary		
Matching words	5,628,746	67.15%
Common nouns	4,195,424	
Noun adjectival verbs	532,336	
Nouns anomalous conjunction of the "sa" series	330,599	
Proper nouns (places)	408,258	
Proper nouns (persons)	34,143	
Proper nouns - general	7,232	
Proper nouns (organizations)	119,994	
Numerals	534	
Adverbs	226	

3.2 Analysis of results

The following table lists, from among the non-matching data, 30 frequently used three-character compounds from each company's dictionary.

<table>
<tr><th colspan="3">Company A non-matching data and frequency</th><th colspan="3">Company B non-matching data and frequency</th></tr>
<tr><td>001</td><td>Ej;q2H</td><td>19386</td><td>001</td><td>6eFsG/</td><td>13565</td></tr>
<tr><td>002</td><td>==8'F|</td><td>15425</td><td>002</td><td>6e0IG/</td><td>12852</td></tr>
<tr><td>003</td><td>==H,F|</td><td>15197</td><td>003</td><td>H,6eG/</td><td>12109</td></tr>
<tr><td>004</td><td>==6eF|</td><td>15175</td><td>004</td><td>7nO"B3</td><td>11538</td></tr>
<tr><td>005</td><td>==<7F|</td><td>14838</td><td>005</td><td>6e;OG/</td><td>10000</td></tr>
<tr><td>006</td><td>==O;F|</td><td>14743</td><td>006</td><td>JR@/I\</td><td>9640</td></tr>
<tr><td>007</td><td>==FsF|</td><td>14334</td><td>007</td><td>F|8a8e</td><td>8952</td></tr>
<tr><td>008</td><td>==;MF|</td><td>13714</td><td>008</td><td>H,H,G/</td><td>8900</td></tr>
<tr><td>009</td><td>==;OF|</td><td>13601</td><td>009</td><td>:IG/Kv</td><td>8579</td></tr>
<tr><td>010</td><td>6eFsG/</td><td>13565</td><td>010</td><td>LL;2>H</td><td>8217</td></tr>
<tr><td>011</td><td>==0IF|</td><td>13316</td><td>011</td><td>F|8aA0</td><td>7408</td></tr>
<tr><td>012</td><td>6e0IG/</td><td>12852</td><td>012</td><td>H,<7G/</td><td>6661</td></tr>
</table>

3.3 Comparison of the results

While the Company A dictionary is very good, it does not appear to have analyzed the data compiled from the volumes of Nikkei corpora. It is also quite conspicuous that the dictionary does not contain such terms as "investor" "Ej;q2H" and "high level," "9b?e=^" nor does it handle numerals adequately. It also needs to include place names and the like.

Company B was able to achieve better results probably because as a private company producing word processors and other products, it would have conducted a broad-ranging and extensive examination. And I believe these days it has produced an even better dictionary. I shall leave the decision on which is the better dictionary up to those who read this paper.

3.4 Collation of borrowed words

I collated the borrowed words (words of foreign origin written in katakana) in the two companies' dictionaries with five years' worth of Nihon Keizai Shimbun data.

Borrowed words in Nihon Keizai Shimbun data

	Distinct words	Total occurrences	
	194,576	7,826,426	100.00%
Matching in both dictionaries	6,323	5,452,876	69.67%
Matching only in Company A dictionary	8,904	373,679	4.77%
Matching only in Company B dictionary	3,330	695,120	8.88%

Of the Nikkei data, 74.44% is matched with Company A's dictionary, and 78.55% with Company B's. There are 9,653 matching words in Company B's dictionary and 15,227 in Company A's. The reason for

this is probably that Company B used frequency of use better than Company A when selecting words. Generally Company A seems to have chosen borrowed words whose usage is well established and focused less on current words, mainly because it used dictionaries designed for people when selecting the borrowed word entries.

3.5 Headwords and Japanese reading of kanji

The Japanese reading of the kanji are attached as kana to the headwords, and while they generally match, there are some differences, as the following shows.

(1) Differences in reading of two-character compounds from Nikkei corpora

	Number of cases	Total words
Japanese reading is different from first character	273	33,438
One of the readings has an extra character (difference of one character between the two readings in the number of distinct characters)	84	140,423
One of the characters in the two readings is different (making the number of different characters two; e.g. one reading uses a $8 and the other a $B)	10	1,396
One of the characters in the two readings is different (different from the above example; making the number of different characters two)	71	5,989
Two of the characters in the two readings are different (making the number of different characters three or four)	46	7,139
Three of the characters in the two readings are different (making the number of different characters five or more)	15	10,332
Multiple readings are given, so there will be differences in readings and also variations in the number of readings	246	141,257
	745	339,974

(2) Differences in reading of three-character compounds from Nikkei corpora

	Number of cases	Total words
Japanese reading is different from first character	13	192
One of the readings has an extra character (difference of one character between the two readings in the number of distinct characters)	5	3,978
One of the characters in the two readings is different (making the number of different characters two; e.g. one reading uses a $8 and the other a $B)	3	124
One of the characters in the two readings is different (different from the above example; making the number of different characters two)	18	1,755
Two of the characters in the two readings are different (making the number of different characters three or four)	19	1,390
Three of the characters in the two readings are different (making the number of different characters five or more)	5	203
Multiple readings are given, so there will be differences in readings and also variations in the number of readings	5	79
	68	7,721

These differences can be put down to either differences in attaching kana readings to the different parts of speech resulting in different kana being used to show declension or inflection, or the differences in classical and contemporary kana renderings of certain compounds.

A check of the readings that did not match showed that Company A's dictionary tended to use the older readings and Company B's dictionary the more contemporary readings.

For example, Company A used the reading ooakindo for the compound Bg>&?M(major trader), while Company B used daishonin. It is only natural that differences such as these will appear in large volumes of data, but even so, it is hoped that we can gradually cut them out as we progress.

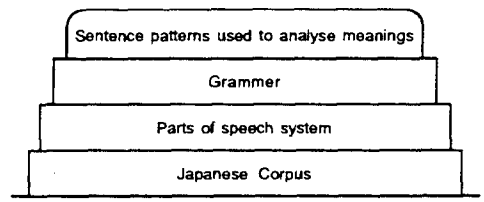
3.6 Headwords and parts of speech

While there are differences between the two dictionaries in their system of parts of speech, they are only minor and can be disregarded using the following matrix.

Company B (Two-character compounds) Parts of speech	Company A (Two-character compounds) Parts of speech	Correlative particles	Temporal adverbs	Adverbs	Common nouns	Proper nouns	Total	Total data
Proper nouns-general								
Noun adjectival verbs								
Numerals								
Noun anomalous conjigation of the "za" series								
Proper nouns(places)								
Adverbs								
Proper nouns(persons)								
Common nouns								
Proper nouns(organizations)								
Noun anomalous conjugation of the "za" series								
Total								
Total data								

I originally thought I would not have to worry about minor differences in the system, but I realized the opposite was the case when I found that different parts of speech were used in different places. The differences were in fact a substantial problem. Company B's dictionary is a machine-readable dictionary designed for kana and kanji characters, but it can also be used freely for natural language processing.

Grammar is based on the parts of speech system, so variations in the system will natural cause variations in the grammar itself. Without a correct parts of speech system, correct grammar is not possible.



4 Short and Long Terms as Headwords

Whether to use long terms or short terms as headwords is an important decision. Short terms are mainly used in the following cases.

1. In applied fields such as information retrieval
2. For dictionaries used by people
   i) to improve the collation rate
   ii) to expand the content and increase headwords without increasing the number of dictionaries and the thickness

   In i) and ii), short headwords are used because people analyze the results and understand the meaning of the words.

Long terms are mainly used in the following cases.

1. In applied fields such as machine translation
2. Natural language processing where meanings have to be processed

Long terms are used for technical terminology. Although effective for removing ambiguity, long terms do have their drawback in that they increase the volume of the dictionary and lower the collation rate (appearance rate). Therefore perseverance is necessary when collecting these terms. Cutting long terms down to short terms with the structure intact is, however, quite simple using the following method.

Natural language processing → ((natural language) processing) ← Term can be written with a space between the three distinct words while keeping the structure intact

There is, however, no one single accepted way of converting the three distinct words back into the one term. Several combinations are possible, and from among these, it is difficult to identify a common rule for determining which should be used.

5 Future Issues

1. We have to look at how to collect headwords, and what should be included in the dictionaries. This is a quality-related issue as well.
2. In using machine-readable dictionaries, we have to know how the dictionary was developed, what its special features are, how it functions, and its capability. These aspects are rarely, if ever, indicated in the dictionary. A well balanced assessment on the quality of a dictionary cannot be made just on the number of headwords.
3. We have to collate machine-readable dictionaries that have been designed for people to look for words, read and understand and dictionaries designed for machine processing to determine what is missing and therefore what has to be added.

6 Conclusion

I have realized that while selecting headwords is a simple process, it is nonetheless a vital part of developing a dictionary. And although computers are essential for analyzing word frequency within corpora, even a small human input can achieve significant results.

It is clear where priority should be placed, and this will reduce the effort used in making the necessary judgements. Unless corpora are analyzed regularly, past analysis results will quickly become outdated. We have to systematize the experience gained by our predecessors in this field, and work out the most effective way of using this bank of experience in computer processing. I have also found out that we need to examine the dictionary headwords and content rather than just rely on computer analysis of large volumes of corpora.

Several companies are producing machine-readable dictionaries on CD-ROM, and I believe we need to refer to these, together with our analysis of corpora, in our aim of developing an extensive machine-readable dictionary. It would be preferable to develop a dictionary with roughly one million headwords.

Similar issues arise in the development of technical dictionaries, so it is necessary to research and build up corpora in the technical fields.

I am confident that the quality and capability of the dictionaries will only be improved as more and more people use them and their suggestions and criticisms are taken on board by the developers and manufacturers.

Regarding the data

Data used in this analysis are from Nihon Keizai Shimbun CD-ROM, 1990, 1991, 1992, 1993 and 1994 editions, purchased from Nikkei Sogo Hanbai Co., Ltd.

References

[1] K. M. Elizabeth Murray, translated by Tomomi Kato, Kotoba e no Jonetsu, Pts 1 and 2, Sanseido, May 1984.

[2] Takenori Kenbo, Jisho wo Tsukuru, Tamagawa Sensho, Nov. 1976.

[3] Satomi Nishiyama & QQQ, Jisho ga Konna ni Omoshirokute Ii Kashira, JICC, June 1996

[4] Yasuhito Tanaka, Update of Machine-Readable Dictionaries Natural Language Processing, 112-17, Information Processing Institute, March 1996.