Ranking Words for Building a Japanese Defining Vocabulary

Tomoya Noro
Department of Computer Science
Tokyo Institute of Technology
2-12-1 Meguro, Tokyo, 152-8552 Japan
noro@tt.cs.titech.ac.jp

Takehiro Tokuda
Department of Computer Science
Tokyo Institute of Technology
2-12-1 Meguro, Tokyo, 152-8552 Japan
tokuda@cs.titech.ac.jp

Abstract

Defining all words in a Japanese dictionary by using a limited number of words (defining vocabulary) is helpful for Japanese children and second-language learners of Japanese. Although some English dictionaries have their own defining vocabulary, no Japanese dictionary has such vocabulary as of yet. As the first step toward building a Japanese defining vocabulary, we ranked Japanese words based on a graph-based method. In this paper, we introduce the method, and show some evaluation results of applying the method to an existing Japanese dictionary.

1 Introduction

Defining all words in a dictionary by using a limited number of words (defining vocabulary) is helpful in language learning. For example, it would make it easy for children and second-language learners to understand definitions of all words in the dictionary if they understand all words in the defining vocabulary. In some English dictionaries such as the Longman Dictionary of Contemporary English (LDOCE) (Proctor, 2005) and the Oxford Advanced Learner's Dictionary (OALD) (Hornby and Ashby, 2005), 2,000-3,000 words are chosen and all headwords are defined by using the vocabulary. Such dictionaries are widely used for language learning.

Currently, however, such a dictionary in which a defining vocabulary is specified has not been available in Japanese. Although many studies for Japanese "basic vocabulary" have been done (National Institute for Japanese Language, 2000), "basic vocabulary" in the studies means a vocabulary which children or second-language learners have (or should learn). In other words, the aim of such studies is to determine a set of headwords which should be included in a Japanese dictionary for children or second-language learners.

We think that there is a difference between "defining vocabulary" and "basic vocabulary". Although basic vocabulary is usually intended for learning expression in newspaper/magazine articles, daily conversation, school textbook, etc, a defining vocabulary is intended for describing word definition in a dictionary. Some words (or phrases) which are often used in word definition, such as "... の略 (abbreviation of ...)", "転じて (change/shift)" [1], "物事 (thing/matter)" etc, are not included in some kinds of basic vocabulary. Additionally only one word in a set of synonyms should be included in a defining vocabulary even if all of them are well-known. For example, if a word "使う (use)" is included in a defining vocabulary, synonyms of the word, such as "使用する", "利用する" and "用いる", are not needed.

A goal of this study is to try to build a Japanese defining vocabulary on the basis of distribution of words used in word definition in an existing Japanese dictionary. In this paper, as the first step of this, we introduce the method for ranking Japanese words, and show some evaluation results of applying the method to an existing Japanese dictionary. Also, we compare the results with two kinds of basic vo-

[1] It is a kind of conjunction used to describe a new meaning comes out of the original meaning.

Headword	Word definition
硬貨 (kouka)	金属製の貨幣。
紙幣 (shihei)	金属貨幣の代用として流通する紙の貨幣。
外貨 (gaika)	外国の貨幣。
贋金 (nisegane)	にせの貨幣（特に、硬貨）。

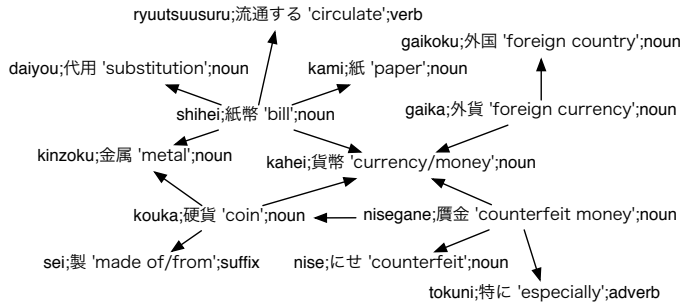

Figure 1: A word reference graph

cabulary, and discuss the difference.

2 Related Work

Kasahara et al. constructed a Japanese semantic lexicon, called "Lexeed" (Kasahara et al., 2004). The lexicon contains the most familiar 28,000 Japanese words, which are determined through questionnaires. All words in the lexicon are defined by using 16,900 words in the same lexicon. However, the size of the vocabulary seems to be too large compared to the size of the defining vocabularies used in LDOCE and OALD. We also think that whether a word is familiar or not does not always correspond to whether the word is necessary for word definition or not.

Gelbukh et al. proposed a method for detecting cycles in word definitions and selecting primitive words (Gelbukh and Sidorov, 2002). This method is intended for converting an existing "human-oriented" dictionary into a "computer-oriented" dictionary, and the primitive words are supposed not to be defined in the dictionary.

Fukuda et al. adopted an LSA-based (latent semantic analysis) method to build a defining vocabulary (Fukuda et al., 2006). The method would be another solution to this issue although only a small evaluation experiment was carried out.

3 Method

Our method for building a Japanese defining vocabulary is as follows:

1. For each headword in an existing Japanese dictionary, represent the relationship between the headword and each word in the word definition as a directed graph (word reference graph).
2. Compute the score for each word based on the word reference graph.
3. Nominate the high ranked words for the Japanese defining vocabulary.
4. Manually check whether each nominated word is appropriate as defining vocabulary or not, and remove the word if it is not appropriate.

In the rest of this section, we introduce our method for constructing word reference graph and computing score for each word.

3.1 Word Reference Graph

A word reference graph is a directed graph representing relation between words. For each headword in a dictionary, it is connected to each word in the word definition by a directed edge (Figure 1). Nodes in the graph are identified by reading, base form (orthography), and parts-of-speech because some words have more than one part-of-speech or reading ("余り (the reading is '*amari*')" has two parts-of-speech, noun and adverb, and "小節" has two readings, "*shousetsu*" and "*kobushi*"). Postpositions, auxiliary verbs, numbers, proper names, and symbols are removed from the graph.

3.2 Computing The Score for Each Word

The score of each word is computed under the assumption that

1. A score of a word which appears in many word definitions will be high.
2. A score of a word which appears in the definition of a word with high score will also be high.

If a word is included in a defining vocabulary, words in the word definition may need to be included in order to define the word. The second assumption reflects the intuition. We adopt the algorithm of PageRank (Page et al., 1998) or LexRank (Erkan and Radev, 2004), which computes the left eigenvector of the adjacency matrix of the word reference graph with the corresponding eigenvalue of 1.

4 Evaluation

4.1 Experimental Setup

We used the Iwanami Japanese dictionary corpus (Hasida, 2006). The corpus was built by annotating the Iwanami Japanese dictionary (the 5th edition) with the GDA tags (Hasida, 2004) and some other tags specific to the corpus. Although it has many kinds of tags, we focus on information about the headword (hd), orthography (orth), part-of-speech (pos), sentence unit in word definition (su), and morpheme (n, v, ad, etc.). We ignore kind of additional information, such as examples (eg), grammatical explanations (gram), antonyms (ant), etymology (etym), references to other entries (sr), etc, since such information is not exactly "word definition". Words in parentheses, "「」" and "『』", are also ignored since they are used to quote some words or expressions for explanation and should be excluded from consideration of defining vocabulary.

Some problems arose when constructing a word reference graph.

1. **Multiple ways of writing in *kanji*:**

   For example, in the Iwanami Japanese dictionary, "引く", "弾く", "曳く", "牽く", "碾く", "轢く" and "退く" appear in an entry of a verb "*hiku*" as its orthography. If more than one writing way appear in one entry, they are merged into one node in the word reference graph (they are separated if they have different part-of-speech).

2. **Part-of-speech conversion:**

   While each word in word definition was annotated with part-of-speech by corpus annotators, part-of-speech of each headword in the dictionary was determined by dictionary editors. The two part-of-speech systems are different from each other. In order to resolve the difference, we prepared a coarse-grained part-of-speech system (just classifying into noun, verb, adjectives, etc.), and converted part-of-speech of each word.

3. **Word segmentation:**

   In Japanese, words are not segmented by spaces and the word segmentation policy for corpus annotation sometimes disagree with the policy for headword registration of the Japanese Iwanami dictionary. In the case that two consecutive nouns or verbs are in word definition and a word consisting of the two words is included as a headword in the dictionary, the two words are merged into one word.

4. **Difference in writing way between a headword and a word in word definition:**

   In Japanese language, we have three kind of characters, *kanji*, *hiragana*, and *katakana*. Most of the headwords appearing in a dictionary (except loanwords) are written in *kanji* as orthography. On the other hand, for example, "事 (matter)" is usually written in *hiragana* ("こと") in word definition. However, it is difficult to know automatically that a word "こと" in word definition means "事", since the dictionary has other entries which has the same reading "*koto*", such as "琴 (Japanese harp)" and "古都 (ancient city)". We merged two nodes in the word reference graph manually if the two words are the same and only different in the writing way.

As a result, we constructed a word reference graph consisting of 69,013 nodes.

We adopted the same method as (Erkan and Radev, 2004) for computing the eigenvector of the adjacency matrix (score of each word). Damping factor for random walk and error tolerance are set to 0.15 and 10^{-4} respectively.

4.2 Result

Table 1 shows the top-50 words ranked by our method. Scores are normalized so that the score of the top word is 1.

Table 1: The top-50 words

	Score	Reading	Orthography	POS	Meaning
1	1.000	*aru*	有る, 在る	V	exist
2	.7023	*i*	意	N	meaning
3	.6274	*aru*	或る	Adn *	certain/some
4	.5927	*koto*	事	N	matter
5	.5315	*suru*	為る	V	do
6	.3305	*mono*	物, 者	N	thing/person
7	.2400	*sono*	其の	Adn *	its
8	.2118	*hou*	方	N	direction
9	.1754	*tatsu*	立つ, 建つ	V	stand/build
10	.1719	*mata*	又, 復, 亦	Conj	and/or
11	.1713	*iru*	居る, 処る	V	exist
12	.1668	*hito*	人	N	person
13	.1664	*tsukau*	使う, 遣う	V	use
14	.1337	*iku*	行く, 往く	V	go/die
15	.1333	*naru*	成る, 為る 生る	V	become
16	.1324	*iu*	言う, 云う 謂う	V	say
17	.1244	*monogoto*	物事	N	thing/matter
18	.1191	*dou*	同	Adn *	same
19	.1116	*sore*	其れ	Pron	it
20	.1079	*toki*	時, 刻	N	time
21	.1074	*teki*	的	Suffix	-like
22	.1020	*souiu*	そういう	Adn *	such
23	.09682	*joutai*	状態	N	situation
24	.09165	*arawasu*	表す, 現す 顕す, 著す	V	represent/ appear/ write a book
25	.08968	*ieru*	言える	V	can say
26	.08780	*ei*	A	N	A
27	.08585	*ten*	点	N	point
28	.08526	*tokuni*	特に	Adv	especially
29	.08491	*go*	語	N	word
30	.08449	*iiarawasu*	言い表す	V	express
31	.08255	*matawa*	又は	Conj	or
32	.07285	*erabitoru*	選び取る	V	choose & take
33	.07053	*baai*	場合	N	case
34	.06975	*tokoro*	所, 処	N	place
35	.06920	*katachi*	形	N	shape
36	.06873	*nai*	無い	Adj	no
37	.06855	*kotogara*	事柄	N	matter
38	.06709	*bii*	B	N	B
39	.06507	*yakunitatsu*	役に立つ	V	useful
40	.06227	*wareware*	我我	Pron	we
41	.06109	*joshi*	助詞	N	postposition
42	.06089	*iitsukeru*	言いつける	V	tell
43	.06079	*ten*	転	N	change/shift
44	.05989	*eigo*	英語	N	English language
45	.05972	*jibun*	自分	N	self
46	.05888	*kata*	方	Suffix	way
47	.05879	*tame*	為	N	reason/aim
48	.05858	*kaku*	書く, 描く	V	write/draw/ paint
49	.05794	*kangaeru*	考える 勘える	V	think
50	.05530	*fukushi*	副詞	N	adverb

* "Adn" indicates "adnominal word", which is a Japanese-specific category and always modifies nouns.

From the result, we can find that not only common words which may be included in a "basic vocabulary", such as "有る (exist)", "或る (certain/some)" [2], "為る (do')", "物 (thing)", etc., but also words which are not so common but are often used in word definition, such as "意 (meaning)", "物事 (thing/matter)", "転 (change/shift)".

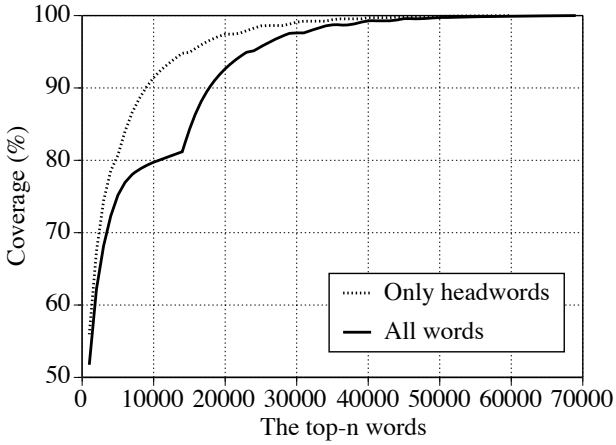

Figure 2: Word coverage

On the other hand, some words in the top ranked words, such as "A" and "B", seem not to be appropriate for defining vocabulary. These words appear only in word definition and are not included in the Iwanami Japanese dictionary as headwords (i.e. unregistered words) [3]. The score of an unregistered word tends to be higher than it should be, since the node corresponding to the word has no edge to other nodes in the word reference graph.

Figure 2 shows word coverage, i.e. percentage of words appearing in word definition which were ranked in the top-n. From the result (solid line), we can find that the increase in coverage around $n = 10,000$ is low and the coverage increases suddenly from $n = 15,000$. This is because all unregistered words were ranked in the top-15000. If all unregistered words are removed, the increase in coverage gets gradually lower as n increases (dotted line).

In construction of a word reference graph, 9,327 words were judged as unregistered words. The reason is as follows:

1. **Part-of-speech mismatch:**

   In order to solve the difference between the part-of-speech system for annotation of headwords and the system for annotation of words in the definition of each headword, we pre-

[2] It is used to say something undetermined or to avoid saying something exactly even if you know that.

[3] In some word definitions, roman letters are used as variables.

pared a coarse-grained part-of-speech system and converted part-of-speech of each word. However, the conversion failed in some cases. For example, some words are annotated with suffix or prefix in word definition, while they are registered as noun in the dictionary.

2. **Mismatch of word segmentation:**

   Two consecutive nouns or verbs in word definition were merged into one word if a word consisting of the two words is included as a headword in the Iwanami Japanese dictionary. However, in the case that a compound word is treated as one word in word definition and the word is not registered as a headword in Iwanami Japanese dictionary, the word is judged as an unregistered word.

3. **Error in format or annotation of the corpus:**

   Since the Iwanami Japanese dictionary corpus has some errors in format or annotation, we removed entries which have such errors before construction of the word reference graph. Headwords which were removed for this reason are judged as unregistered words.

4. **Real unregistered words:**

   Some words in word definition are not registered as headwords actually. For example, although a noun "英語 (English language)" appears in word definition, the word is not registered as a headword.

Unregistered words should carefully be checked whether they are appropriate as defining vocabulary or not at the third step of our method described in section 3.

4.3 Comparison

In order to look at the difference between the result and so-called "basic vocabulary", we compared the result with two types of basic vocabulary: one was built by the National Institute for Japanese Language (including 6,099 words) and the other was built by the Chuo Institute for Educational Research (including 4,332 words) (National Institute for Japanese Language, 2001). These two types of vocabulary are intended for foreigners (second-language learners) and Japanese children (elementary school students) respectively.

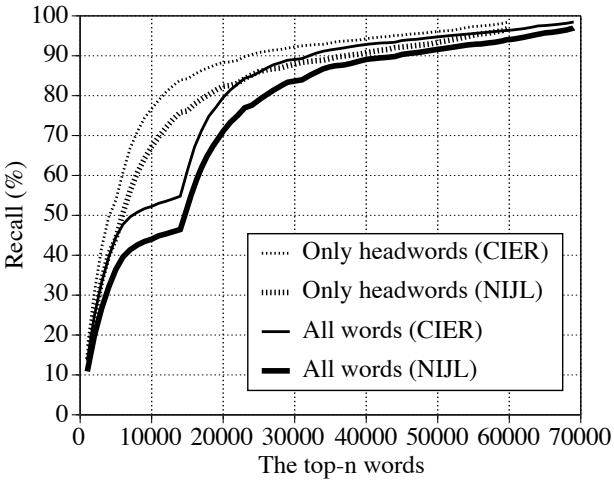

Figure 3: Comparison with two types of basic vocabulary

Table 2: High-ranked words out of the two basic vocabularies

Rank	Reading	Orthography	POS	Meaning
51	*tenjiru*	転じる	V	shift/change
102	*youhou*	用法	N	usage
113	*ryaku*	略	N	abbreviation
372	*furumai*	振舞い	N	behavior
480	*sashishimesu*	指し示す	V	indicate

Figure 3 shows recall, i.e. percentage of the number of words appearing in both our result and each vocabulary out of the number of words in the vocabulary. As in the case of word coverage, the increase in recall around $n = 10,000$ is low if unregistered words are not removed (solid lines). If the same number of headwords as the size of each basic vocabulary are picked up from our result, it can be found that about 50% of the words are shared with each basic vocabulary (dotted lines).

Some of the high-ranked words out of the two basic vocabularies and some of the low-ranked words in the vocabularies are listed in Table 2 and 3. Although it would be natural that the words listed in Table 2 are not included in the basic vocabularies, they are necessary for describing word definition. On the other hand, the words listed in Table 3 may not be necessary for describing word definition, while they are often used in daily life.

Table 3: Low-ranked words in the two basic vocabularies

Rank	Reading	Orthography	POS	Meaning
20062	*taifuu*	台風	N	typhoon
20095	*obaasan*	お婆さん	N	grandmother
31097	*tetsudau*	手伝う	V	help/assist
37796	*kamu*	噛む	V	bite
47579	*mochiron*	勿論	Adv	of course
65413	*tokoroga*	ところが	Conj	but/however

5 Conclusion

In this paper, we introduced the method for ranking Japanese words in order to build a Japanese defining vocabulary. We do not think that a set of the top-n words ranked by our method could be a defining vocabulary as is. The high ranked words need to be checked whether they are appropriate as defining vocabulary or not.

As described in section 1, defining all words with a defining vocabulary is helpful in language learning. In addition, we expect that the style of writing word definitions (e.g. which word should be used, whether the word should be written in *kanji* or *hiragana*, etc.) can be controlled with the vocabulary.

This kind of vocabulary could also be useful for NLP researches as well as language learning. Actually, defining vocabularies used in LDOCE and OALD are often used in some NLP researches.

The future work is the following:

- The size of a defining vocabulary needs to be determined. Although all words in LDOCE or OALD are defined by 2,000-3,000 words, the size of a Japanese defining vocabulary may be larger than English ones.
- Wierzbicka presented the notion of conceptual primitives (Wierzbicka, 1996). We need to look into our result from a linguistic point of view, and to discuss the relation.
- It is necessary to consider how to describe word definition as well as which word should be used for word definition. Definition of each word in a dictionary includes many kinds of information, not only the word sense but also historical background, grammatical issue, etc. Only word sense should be described with a defining vocabulary, since the other information is a little different from word sense and it may be difficult to describe the information with the same vocabulary.

References

Güneş Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Muhtar Fukuda, Yasuhiro Ogawa, and Katsuhiko Toyama. 2006. Automatic generation of dictionary definition words based on latent semantic analysis. In *the 5th Forum on Information Technology*. In Japanese.

Alexander F. Gelbukh and Grigori Sidorov. 2002. Automatic selection of defining vocabulary in an explanatory dictionary. In *the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, pages 300–303.

Koiti Hasida, 2004. *The GDA Tag Set*. http://i-content.org/GDA/tagset.html.

Koiti Hasida, 2006. *Annotation of the Iwanami Japanese Dictionary – Anaphora, Coreference And Argument Structure –*. http://www.i-content.org/rwcDB/iwanami/doc/tag.html (In Japanese).

A. S. Hornby and Michael Ashby, editors. 2005. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press.

Kaname Kasahara, Hiroshi Sato, Francis Bond, Takaaki Tanaka, Sanae Fujita, Tomoko Kanasugi, and Shigeaki Amano. 2004. Construction of Japanese Semantic Lexicon: Lexeed. In *IPSJ SIGNL 159*, pages 75–82. In Japanese.

The National Institute for Japanese Language, editor. 2000. *Japanese Basic Vocabulary – An Annotated Bibliography And a Study –*. Meiji Shoin. In Japanese.

The National Institute for Japanese Language, editor. 2001. *A Basic Study of Basic Vocabulary for Education – Construction of a Database of Basic Vocabulary for Education –*. Meiji Shoin. In Japanese.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University.

Paul Proctor, editor. 2005. *Longman Dictionary of Contemporary English*. Longman.

Anna Wierzbicka. 1996. *Semantics: Primes and Universals*. Oxford University Press.