Building a Database of Japanese Adjective Examples from Special Purpose Web Corpora

Masaya YAMAGUCHI

National Institute for Japanese Language and Linguistics
10-2 Midori-cho, Tachikawa City, Tokyo, Japan
masaya@ninjal.ac.jp

Abstract

It is often difficult to collect many examples for low-frequency words from a single general purpose corpus. In this paper, I present a method of building a database of Japanese adjective examples from special purpose Web corpora (SPW corpora) and investigates the characteristics of examples in the database by comparison with examples that are collected from a general purpose Web corpus (GPW corpus). My proposed method construct a SPW corpus for each adjective considering to collect examples that have the following features: (i) non-bias, (ii) the distribution of examples extracted from every SPW corpus bears much similarity to that of examples extracted from a GPW corpus. The results of experiments shows the following: (i) my proposed method can collect many examples rapidly. The number of examples extracted from SPW corpora is more than 8.0 times (median value) greater than that from the GPW corpus. (ii) the distributions of co-occurrence words for adjectives in the database are similar to those taken from the GPW corpus.

Keywords: Japanese adjective, Web corpus, database of examples

1. Introduction

Analyzing the distribution of co-occurrence words for a word requires many examples of the word. But it is often difficult to collect enough examples for low-frequency words from a single general purpose corpus. An approach to solve this problem is to construct a special purpose corpus for each word and extract examples from the corpus.

This paper presents a method of building a database of Japanese adjective examples from special purpose Web corpora (SPW corpora) and investigates the characteristics of examples in the database by comparison with examples that are collected from a general purpose Web corpus (GPW corpus). Each example in the database has the dependency structure information for the target adjective. Users can search the database by an adjective, and browse examples and the distribution of co-occurrence words for the adjective.

Considering to use the database for linguistic studies, a set of examples for an adjective needs to have the following features: (i) non-bias, (ii) the distribution of examples extracted from a SPW corpus bears much similarity to that of examples extracted from a GPW corpus.

To collect such examples, I construct SPW corpora using a method based on Sharoff (2006), which collects Web pages by random queries to a search engine. A GPW corpus consists of randomly selected sentences from the collected Web pages, which are collected by the same construction method.

The rest of the paper is structured as follows: In section 2, I review related works. In section 3, I define the target Japanese adjectives, and describe the procedure of building the database. In section 5, I evaluate the database. In section 6, I conclude this paper by summarizing evaluations.

2. Related Work

Since the Web has begun to be used as a corpus for linguistic studies in early 2000's (Kilgarriff and Grefenstette, 2003), various Web corpora have been constructed (Baroni and Kilgarriff, 2006; Imai et al., 2013; Sharoff, 2006; Srdanović Erjavec and Nishina, 2008; Suchomel and Pomikálek, 2012). For example, Japanese Web corpora JpWac (Srdanović Erjavec and Nishina, 2008) and JpTenTen (Suchomel and Pomikálek, 2012), that include about 400 million words and about 10 billion words respectively, are provided as a commercial Web service by a corpus retrieval system SketchEngine (Kilgarriff et al., 2004). Imai et al. (2013) collected Web pages based on BootCaT (Baroni and Bernardini, 2004) and released to the public "Tsukuba Web Corpus" for Japanese language education, that includes about 1.1 billion words.

To collect Web pages, the methods of constructing these corpora use URLs that a Web search engine retrieves by random keywords. Baroni and Kilgarriff (2006), Suchomel and Pomikálek (2012) use the URLs as seeds to crawl the Web. On the other hand, Sharoff (2006) collects Web pages by using only the URLs rather than using them for a Web crawler.

With respect to linguistic evaluations for these methods, their corpora have been investigated by comparison with non-Web corpora (e.g. BNC) in terms of the frequency of words, the overlap of vocabulary and so on (Baroni and Kilgarriff, 2006; Sharoff, 2006). But their corpora are not for special purposes and there have not been enough investigations what kind of examples could be extracted from special purpose corpora.

3. A Database of Japanese Adjective Examples

3.1. Target Japanese Adjectives

In Japanese, there are two types of adjectives: i-adjective and na-adjective[1]. In this paper, the target adjectives for collecting examples are 551 i-adjectives, which are all entry adjectives in the dictionary of Japanese morphological

[1]The basic forms of i-adjective and na-adjective end in the letter 'i' and 'na', respectively.

analyzer JUMAN[2]. The reason why na-adjectives are eliminated is there are various theories as to the identification of the part of the speech for them.
Japanese adjectives have three main usages: (1) predicative, (2) adnominal, (3) adverbial usage.

(1) Taiyo-ha *akarui*
sun-ACC bright
(The sun is bright)

(2) *Akarui* hoshi-ga arawareru
bright star-ACC appear
(A bright star appears)

(3) Hoshi-ga *akaruku* kagayaku
star-ACC bright shine
(A star shines bright)

SPW corpora need to be constructed considering these usages, because the dependency structures and the distributions of co-occurrence words for a target adjective are different by usages.
But the corpus construction method described in section 3.2. can not specify a usage but a word form on construction of a corpus. So two SPW corpora are constructed individually for two word forms (basic form and continuous form) of one adjective. As the above usage (1)-(3), examples of predicative and adnominal usages (Usage1, 2) can be collected from a SPW corpus of basic form, and examples of adverbial usage (Usage3) can be done from a SPW corpus of continuous form. As a result, 1102 SPW corpora are constructed to build a database of Japanese adjective examples.

3.2. Constructing special purpose Web corpora

The procedure of corpus construction in this paper is based on Sharoff (2006). As mentioned in section 2., Sharoff (2006) uses random queries to a Web search engine to collect Web pages. This method is expected to collect less biased Web pages than Baroni and Kilgarriff (2006; Suchomel and Pomikálek (2012), because this method collects Web pages by using only URLs that are retrieved by the random queries, while Baroni and Kilgarriff (2006; Suchomel and Pomikálek (2012) use the URLs as seeds to crawl the Web.
The main difference between my proposed method and Sharoff (2006) is the way of removing duplicated contents. Sharoff (2006) removes them page by page in the process of constructing a corpus. On the other hand, my proposed method does them sentence by sentence in the process of building a database of examples (refer to section 3.3.), because my goal is not to construct a corpus but to collect examples for a given word. The process of a SPW corpus construction is as follows.

1. Retrieving up to 50 URLs through Bing Search API[3]. The keywords to the search engine are a target adjective and a noun that is selected randomly from 3000 high frequency ones.
2. Selecting up to 10 URLs from the search results randomly.
3. Downloading Web pages from the URLs except duplicated pages.
4. Repeating the above process until collecting N Web pages ($N = 2000$ for basic form; $N = 1000$ for continuous form). I decided N to get 500 examples for one target adjective.
5. Arranging the collected Web pages (For example, character code conversion to UTF-8, removal of HTML tags).
6. Annotating morphological information to the text by Japanese morphological analyzer JUMAN.

3.3. Storing examples in the database

After the construction of a SPW corpus for an adjective, target examples are stored in a relational database as follows.

1. Extracting sentences that include the target adjective from the corpus except duplicated sentences.
2. Analyzing every sentence by Japanese dependency and case structure analyzer KNP[4] to get co-occurrence words (case elements and adverbial elements) for the target adjective.
3. Storing each sentence in the database if it has more than one co-occurrence word.

3.4. Searching the database

The database has been released to the public on the Web[5], providing a simple user interface that has the following functions: (i) searching for adjectives, (ii) browsing the distribution of co-occurrence words of an adjective, (iii) listing examples.
Table1, 2 show examples of (ii): the distribution of co-occurrence words for Japanese adjective "wakai [young]" in the adnominal and adverbial usage respectively. Users can select a usage anytime. The value beside a co-occurrence word is the number of pages where the word appears. By clicking a co-occurrence word on a table, Examples of the word are listed.

4. Experiments and Evaluation

To evaluate the database, examples in the database are compared with those that are extracted from a GPW corpus in terms of the number of examples, and the distributions of co-occurrence words for target adjectives.

4.1. Constructing a general purpose corpus

The GPW corpus used in the following experiments consists of randomly selected sentences from Web pages that are collected by using random queries to a Web search engine. The method of collecting Web pages is based on Sharoff (2006).
The procedure of the GPW corpus is as follows.

[2]http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[3]http://datamarket.azure.com/dataset/bing/searchweb
[4]http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[5]http://csd.ninjal.ac.jp/adj/

Table 1: The distribution of co-occurrence words for wakai (adnominal usage)

modified noun		adverbial phrase		subject	
hito [person]	228	yahari[all in all]	13	watasi [I]	16
josei [lady]	181	motto [young**er**]	10	nenrei [age]	8
sedai [generation]	124	sarani [young**er**]	10	hito [person]	6

Table 2: The distribution of co-occurrence words for wakai (adverbial usage)

predicate		adverbial phrase(*)		subject(*)	
mieru [look]	134	tosi [age]	53	hito [person]	19
suru [at a young age]	103	itumademo [forever]	16	tosi [age]	16
mirareru [be looked]	72	tokuni [especially]	7	watasi [I]	12

* : elements for "predicate"

1. Retrieving up to 50 URLs through Bing Search API. The keywords to the search engine are N randomly-selected nouns. In this experiment N is 3. The reason why N is 3 is that in the case of N is 2 same queries may be often created because of the lack the number of combinations of nouns; and queries sometimes yielded no results if N is 4.
2. Selecting up to 10 URLs from the search results randomly.
3. Downloading Web pages from the URLs except duplicated pages.
4. Arranging the downloaded Web pages. (For example, character code conversion to UTF-8, removal of HTML tags).
5. Selecting up to 5 sentences from each Web page randomly.
6. Adding the sentences to the GPW corpus if there has not been the same sentence in the corpus yet.
7. Annotating morphological information to the sentences by Japanese morphological analyzer JUMAN.
8. Repeating the above process until collecting 5.5 million Web pages.

The resulting GPW corpus came from 4 million Web pages and consisted of 3.8 hundred million words; 17 million sentences. In the following experiments, the example extraction from the GPW corpus and the dependency analysis are done by the same way described in section 3.3..

4.2. Number of examples

In this section, extracted examples from SPW corpora are evaluated by comparison with examples that are extracted from the GPW corpus in terms of the number of examples. Figure1, 2, 3 show histograms of the example counts of 551 adjectives (Usage 1, 2, 3). Examples of Usage 1, 2 (Figure1, 2) are extracted from SPW corpora (basic form) and classified based on the dependency structure analysis described in section 3.3.. Filled histograms are results for GPW corpora. non-Filled histograms are results for SPW corpora. The values beside graph legends are median values of examples of adjectives.

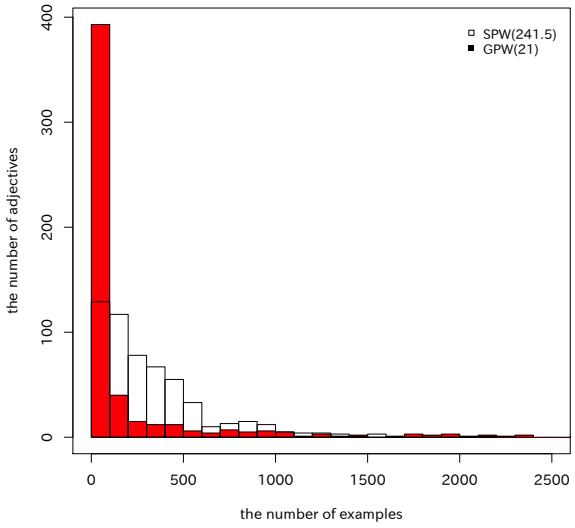

Figure 1: Number of examples (Usage1)

The rates (median value) of extracted examples to collected Web pages are as follows: Usage1 12,1%, Usage2 64.4%, Usage3 64.5%. Examples of Usage1 could not collect the target amount (500 examples) because of the high proportion of examples of Usage2. One strategy to get additional examples is to execute the procedure of section 3.2. repeatedly. The easy extension of a SPW corpus is an advantage of the proposed method. Another strategy is to contrive a better keyword to a search engine (e.g. attaching a period to the end of a target adjective to get examples of Usage1). Since additional example collections for Usage1 were not done in this paper, only Usage2, 3 are evaluated in the pages that follow.

There are also small example adjectives in Usage2, 3. After

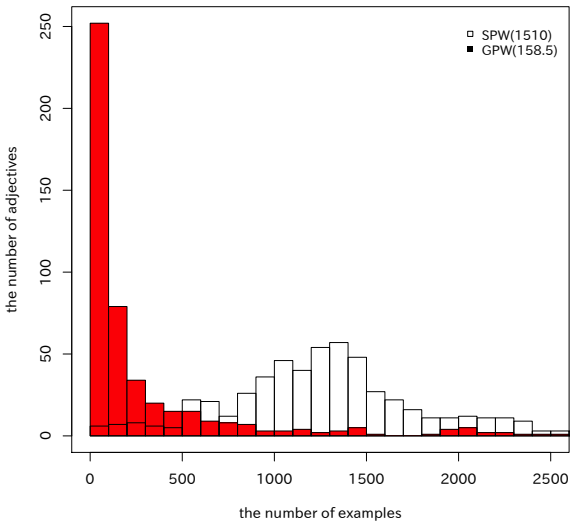

Figure 2: Number of examples (Usage2)

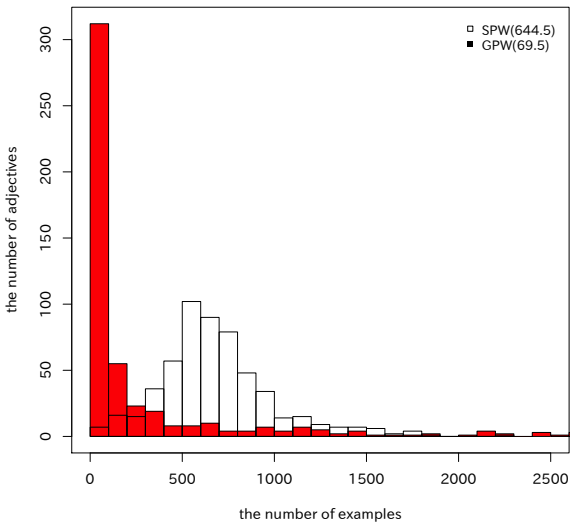

Figure 3: Number of examples (Usage3)

investigating adjectives that have examples of up to 100, these words were found to be classified into two types: (1) shortage of search results (e.g. kotoatarashii), (2) use in a specific word form (e.g. meboshii is used principally in the basic form).

Figure 4 shows two histograms of the ratio of the example counts (SPW corpus / GPW corpus). Note that the scale of the horizontal axis is logarithmic.

These results shows that my proposed method can collect many examples rapidly. For example, the number of examples extracted from SPW corpora of Usage 2 is 9.1 times (median value) greater than that from the GPW corpus. Considering that the GPW corpus came from 5.5 million Web page, 50 million Web pages have to be collected. On the other hand, my proposed method could build the

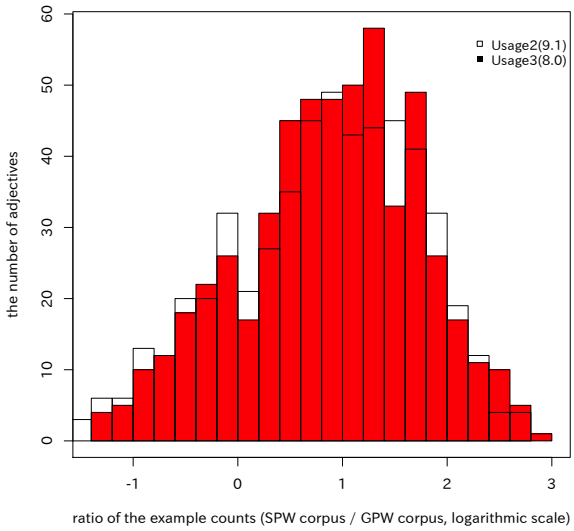

Figure 4: The ratio of the example counts (SPW corpus / GPW corpus)

database by collecting 1.7 million Web pages, and needs only 2000 Web pages to construct a SPW corpus. This rapid construction is an advantage of my proposed method over GPW corpora.

4.3. Similarity of the distribution of co-occurrence words

Examples in the database are evaluated in terms of the similarity between the distributions of co-occurrence words for adjectives in the database and those taken from the GPW corpus. The similarity is defined as cosine similarity $COS(\mathbf{w_1}, \mathbf{w_2})$ as follows:

$$COS(\mathbf{w_1}, \mathbf{w_2}) = \frac{\mathbf{w_1} \cdot \mathbf{w_2}}{|\mathbf{w_1}||\mathbf{w_2}|}$$

where $\mathbf{w_1}$ and $\mathbf{w_2}$ are frequency vectors of co-occurrence words for an adjective in the database and the GPW corpus respectively. An element in a frequency vector is a pair of a co-occurrence word and a label that expresses a syntactic relationship for the target adjective (e.g. [apple, subject] for "the apple is red", [apple, modified noun] for "a red apple"). For avoiding noises, the frequency is treated as zero if the frequency is smaller than 4, and the duplicated element in a Web page is not counted.

Figure 5 shows two histograms of the cosine similarity for 148 adjectives (Usage2) and 127 adjectives (Usage3), that have more than 500 examples.

The median values of cosine similarity of Usage2, 3 are 0.86 and 0.96, respectively. This result proves that the distributions of co-occurrence words for adjectives in the database are similar to those taken from the GPW corpus.

However, some factors to decrease the cosine similarity was found by investigations into the distributions of low cosine similarity adjectives: (i) proper noun phrases (e.g. a title of a movie, Hobbit: omoigakenai[Unexpected]

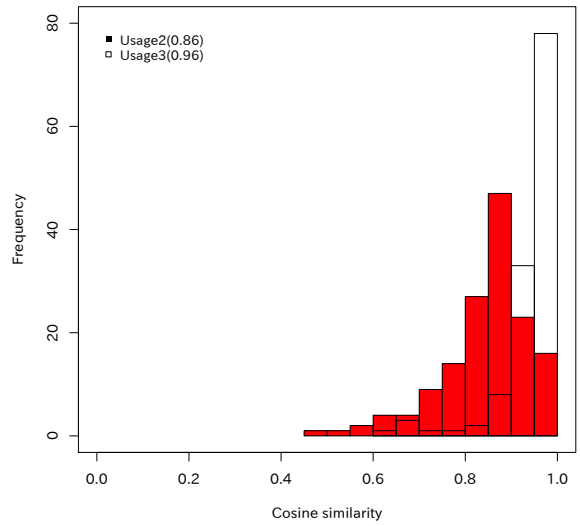

Figure 5: Cosine similarity between the distributions of co-occurrence words in the database and those taken from the GPW corpus

tabi[Journey]), (ii) nouns that express a very general concept (e.g. nagai[long] *aida [time]*). The frequency of (i) in SPW corpora tends to be higher than that in the GPW corpus, while the frequency of (ii) in SPW corpora tends to be lower than that in the GPW corpus. These results are anticipated to be caused by the search engine, because it ranks Web pages that include proper noun phrases like famous movies titles higher.

5. Conclusions

This paper presented a method of building a database of Japanese adjective examples from SPW corpora and evaluated the database by comparison with examples that are collected from the GPW corpus.

The results of experiments showed the following: (i) my proposed method can collect many examples rapidly. The number of examples extracted from SPW corpora is more than 8.0 times (median value) greater than that from the GPW corpus. (ii) the distributions of co-occurrence words for adjectives in the database are similar to those taken from the GPW corpus.

6. References

Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*.

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 87–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shingo Imai, Shiro Akasegawa, and Prashant Pardesh. 2013. Development of NLT: the search tool for Tsukuba Web Corpus. In *Proceedings of the 3rd Workshop of Japanese Corpus Linguistics*, pages 199–206.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Comput. Linguist.*, 29(3):333–347, September.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The sketch engine. In *Proceedings of EURALEX*.

Serge Sharoff. 2006. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11:435–462.

Irena Srdanović Erjavec and Kikuko Nishina. 2008. The Sketch Engine corpus query tool for Japanese and its possible applications. *Japanese Linguistics*, 23:59–80, apr.

Vít Suchomel and Jan Pomikálek. 2012. Efficient web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pages 39–43, Lyon.