

Retrieving Japanese specialized terms and corpora from the World Wide Web

Marco Baroni

SSLMIT, University of Bologna
Corso della Repubblica 136
47100 Forlì, Italy
baroni@sslmit.unibo.it

Motoko Ueyama

SSLMIT, University of Bologna
Corso della Repubblica 136
47100 Forlì, Italy
motoko@sslmit.unibo.it

Abstract

The BootCaT toolkit (Baroni and Bernardini, 2004) is a suite of perl programs implementing a procedure to bootstrap specialized corpora and terms from the web using minimal knowledge sources. In this paper, we report ongoing work in which we apply the BootCaT procedure to a Japanese corpus and term extraction task in the hotel terminology domain. The results of our experiments are very encouraging, indicating that the BootCaT procedure can be successfully applied, with relatively small modifications, to a language very different from English and the other Indo-European languages on which we tested the procedure originally.

1 Introduction

The World Wide Web is a rich source of easily accessible language data (Kilgarriff and Grefenstette, 2003). Among those who can benefit from this resource, there are language professionals (language teachers, translators, interpreters, etc) who routinely work with a variety of specialized languages, where new terms are introduced at a fast pace.

We recently introduced the BootCaT toolkit,¹ a suite of perl programs implementing an iterative knowledge-poor procedure to bootstrap specialized corpora and term lists from the web.

In this paper, we report preliminary results from an ongoing study in which we use the BootCaT tools to extract Japanese hotel business terminology. The study started with practical motivations, that is, the interest of our Italian students of Japanese in this domain and the consequent need to build the relevant language resources for teaching. The study is also giving us a chance to test the cross-linguistic viability of the BootCaT tools by applying the procedure

to a typologically (and orthographically) very different language.

The rest of the paper is structured as follows: In section 2 we shortly review some related work. In section 3 and section 4 we describe the BootCaT procedure and how we tuned it for Japanese, respectively. In section 5 we present our experiments. We conclude in section 6 by sketching some future directions.

2 Related work

The idea of building a corpus using automated search engine queries originates from Ghani et al. (2001), who applied it to the creation of minority language corpora. Our corpus-comparison-based term extraction methodology was inspired by Rayson and Garside (2000). There is, of course, a large body of work on Japanese terminology, some of it involving web mining. For example, Fujii and Ishikawa (2000) use the web to search for definitions of pre-selected terms.

However, as far as we know, this is the first study presenting a full knowledge-poor procedure to extract Japanese terms and specialized corpora from the web.

3 The BootCaT procedure

The main corpus/term bootstrapping loop of the BootCaT procedure is illustrated in Figure 1. The bootstrapping process starts with a small list of seed terms representative of the investigated domain (hotel terminology in the present study). The seeds are randomly combined, and each combination is used as a Google query string. The top n pages returned from each query are retrieved and formatted as text. New seeds are extracted by comparing the frequency of words/terms in the retrieved corpus and in a reference corpus. In the current study, corpus comparison statistics are computed with the UCS toolkit (Evert, 2004). Random combinations

¹BootCaT stands for *Bootstrapping Corpora and Terms*. The toolkit is freely available from:
<http://sslmit.unibo.it/~baroni/bootcat.html>

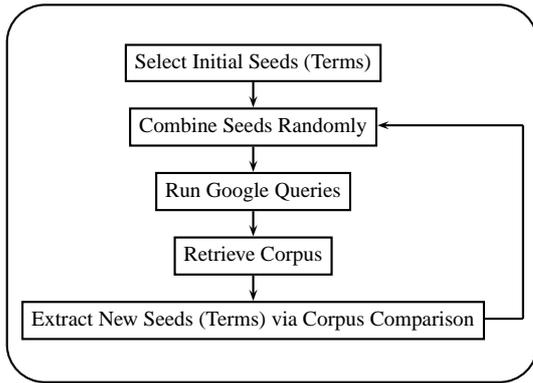


Figure 1: The BootCaT loop

of the newly extracted seed terms are then used for another round of Google queries, and a new corpus is created by retrieving and formatting the pages found in this round. The iterative procedure of terms/corpus extraction can be repeated as many times as desired (e.g., until the corpus reaches a certain size).

4 Adaptation to Japanese

There are two important issues in adapting the procedure described above to Japanese. First, Japanese web-pages can be in different character sets (*shift-jis*, *euc-jp*, *iso-2022-jp*, *utf-8*); second, in Japanese words/tokens are not separated by whitespace or other delimiters. To solve the first problem, we changed the code of the BootCaT script to retrieve and format web-pages. Now, this script detects the character set used to encode a page in the HTML code, and it converts the text of the page from the specified character set into *utf-8*. Since ChaSen (see below) expects input and output to be coded in *euc-jp*, we use the `recode` command line tool² to convert back and forth between *utf-8* and *euc-jp*.

To split the retrieved text into tokens, we use ChaSen (Matsumoto et al., 2000), a powerful command line tool that performs Japanese tokenization, morphological analysis and POS tagging. The parsing and tokenization rules of ChaSen can be modified via parameter files. For our purposes, we added “under-segmenting” rules to preserve two complex templates, i.e., nominal compounds (e.g., *yoyaku-kakunin* ‘reservation-confirmation’; *ryookin-hyoo* ‘rate-chart’) and nouns prefixed by honorific markers (e.g., *go-yoyaku* ‘HONORIFIC-reservation’).

²<http://recode.progiciels-bpi.ca/>

By adding the nominal compound template to the ChaSen parameter file, we capture many candidate complex terms already in the tokenization phase. Thus, at the moment we do not distinguish between a simple and a complex term extraction phase (as we do, instead, when the BootCaT procedure is applied to Western languages). In future work, we would like to explore more sophisticated methods to extract complex terms in Japanese.

5 Experiments

5.1 Preparation of materials

The second author, using her native speaker knowledge and manual web queries, prepared a list of 126 (simple and complex) terms typical of hotel terminology. 20 out of these 126 terms were used as initial seeds for the bootstrapping process: e.g., *yoyaku* ‘reservation’, *kyaku-sitsu* ‘guest room’, *ruumu-saabisu* ‘room service’. The remaining 106 terms are used for recall-oriented evaluation (see section 5.3.1 below).

Our procedure requires the comparison of the retrieved specialized corpora to a reference corpus. Since we did not own a Japanese corpus, we constructed one in the following way. We prepared a set of seeds by randomly selecting 100 words from the basic vocabulary list of an elementary Japanese textbook (Banno et al., 1999). The seeds were combined to form 100 random triplets, and these were used for Google queries. The corpus obtained by downloading and formatting the pages found in this way contains about 3.5M tokens. While, of course, it is not a balanced corpus, it does include texts belonging to a wide variety of topics, genres and styles.

5.2 Procedure

Using the BootCaT tools, we queried Google for 10 randomly constructed triplets of seeds. We retrieved 77 pages, and we tokenized the contents of those pages with ChaSen. We obtained a first corpus of about 100K tokens. We then used the UCS toolkit to find the most typical tokens of this corpus as compared to the reference corpus. In particular, we ranked the terms on the basis of two association measures, log-likelihood ratio and mutual information, computed on contingency tables of occurrences of terms in the specialized and reference corpora. Before computing mutual information, we filtered out terms that occurred less than 10 times in the specialized corpus.

Log-likelihood ratio and mutual information tend to find items at the opposite ends of the frequency scale. For example, at the top of the list ranked by log-likelihood ratio, we see frequent terms such as *hoteru* ‘hotel’ and *choushoku* ‘breakfast’; at the top of the list ranked by mutual information we see rarer terms such as *karaoke-ruumu* ‘karaoke room’ and *yoyaku-kin* ‘reservation fee’.

Combining the top 100 terms from the log-likelihood ratio and mutual information lists, we obtained a new set of 164 seed terms for the next run. In the second and third runs of the procedure, we built 50 triplets to be used as Google query strings. In the second run, we retrieved 236 pages which, again, we tokenized with ChaSen. The resulting corpus contained about 390K tokens. A new list of terms was extracted with the same corpus comparison method described above. This time, the combined list contained 194 terms. In the third run, we retrieved 225 pages, 865K tokens and 196 combined terms. In total, we retrieved 424 distinct terms. We decided to stop and analyze the data we collected up to this point.

5.3 Evaluation

5.3.1 Term quality

The second author rated all the extracted terms using a 3-point scale: irrelevant terms, somewhat relevant terms, completely relevant terms. The “somewhat relevant” category included toponyms and terms of closely related domains (e.g., travel and transportations). The results of this evaluation are summarized in table 1.

	<i>not relevant</i>	<i>somewhat relevant</i>	<i>very relevant</i>	<i>total terms</i>
1st run, ll	13%	12%	75%	100
1st run, mi	7%	23%	70%	100
1st run, ll+mi	10.9%	16.4%	72.5%	164
2nd run, ll	18%	7%	75%	100
2nd run, mi	15%	25%	60%	100
2nd run, ll+mi	16.4%	16.4%	67%	194
3d run, ll	23%	19%	58%	100
3d run, mi	24%	30%	46%	100
3d run, ll+mi	23.9%	25%	51%	196
combined, ll	16.9%	15.5%	67.4%	212
combined, mi	16.7%	28.2%	54.9%	262
combined, ll+mi	18.1%	23.3%	58.4%	424

Table 1: Relevance of retrieved terms

The results reported in this table are very promising: in the final combined list, almost 60% of the retrieved terms are very relevant, and less than 20%

are completely irrelevant.³

A closer examination of the irrelevant items shows that most of them are grammatical morphemes/words (adverbial suffixes, conjunctions, conjugation endings, etc).⁴ This is particularly true in the log-likelihood lists, since grammatical morphemes tend to be high frequency items. Specifically, the most common grammatical elements extracted by the algorithm are those that are typical of interrogative/exhortative sentences in the polite register (for example, *kudasai* ‘please’). It is not surprising to find a high occurrence of such forms in pages addressed to tourists and potential hotel costumers. Indeed, it may be useful to our target users (teachers and students of specialized languages) to be aware that the language of tourism is rich in this kind of expressions.

We also performed recall-oriented evaluation by counting how many of the 106 non-seed items in our original list of manually picked terms (see section 5.1 above) were ranked by the automated procedure in the top 100/200 terms according to at least one measure. The results are reported in table 2.

	<i>proportion of retrieved pre-selected terms</i>	
	<i>top 100 cutoff</i>	<i>top 200 cutoff</i>
1st run, ll	15%	24.5%
1st run, mi	4.7%	16.9%
1st run, ll+mi	17.9%	26.4%
2nd run, ll	16.9%	26.4%
2nd run, mi	1.8%	4.7%
2nd run, ll+mi	17.9%	30.1%
3d run, ll	6.6%	12.2%
3d run, mi	1.8%	1.8%
3d run, ll+mi	8.4%	14.1%
combined, ll	21.6%	32%
combined, mi	6.6%	19.8%
combined, ll+mi	24.5%	36.7%

Table 2: Recall of pre-selected terms

Even with the maximum recall setting (combined runs and measures, top 200 lists), just above one third of the manually selected terms were retrieved automatically. This is not necessarily bad, in light of our good precision results. It rather seems to sug-

³If we select and combine the top 200 terms found with each measure and on each run, we obtain a total of 752 terms, 21.4% of which irrelevant, 25.9% somewhat relevant and 52.6% very relevant.

⁴In an agglutinative language like Japanese, it is often hard to decide which elements should be considered independent function words and which elements should be treated as grammatical affixes.

gest that the types of terms discovered by the algorithm tend to be complementary to those obtained on the basis of intuition. Interestingly, recall is decidedly lower in the mutual information lists than in the log-likelihood lists. This is probably due to the fact that mutual information is mostly picking up low frequency terms, whereas humans are more inclined to select high frequency terms as representative of a domain.

Looking at the manually selected terms that were not in our final set, first of all we notice that some terms were missed since they are typical of Western hotels (e.g., *nakaniwa* ‘court yard’), whereas the large majority of pages we retrieved pertain to Japanese hotels. Many terms are not present in the tokenized corpus because of segmentation issues. For example, the complex term *yotsuboshi-hoteru* ‘four star + hotel’ was incorrectly analyzed as *yotsuhoshihoteru* = ‘four + star hotel’. Single terms such as *basu* ‘bath’ are often found only as part of (highly ranked) complex terms such as *basu-taoru* ‘bath towel’. For some missed terms, we found their equivalents prefixed by a honorific marker: e.g., *go-yoyaku-torikeshi* instead of *yoyaku-torikeshi* ‘reservation cancellation’. As we said, hotel sites tend to use a polite register, which is partly reflected in the frequent prefixation of the honorific marker *go-*.

5.3.2 Corpus quality

The retrieved corpora are used for term extraction, but they also constitute an important deliverable by themselves. To evaluate the quality of the corpora, we randomly selected 90 downloaded pages (30 pages from each of the three rounds). The second author judged these pages on a 3-point scale, assigning the highest score to pages that are highly informative, very reliable, and completely relevant.

Out of the 30 web-pages selected from the first corpus, 27 pages were assigned the highest rating, 1 page was assigned the intermediate rating, and 2 pages were assigned the lowest rating. Of the 30 web-pages selected from the second corpus, 25 pages were assigned the highest rating, 3 pages were assigned the intermediate rating and 2 pages were assigned the lowest rating. Of the 30 web-pages selected from the third corpus, 24 pages were assigned the highest rating, 2 pages were assigned the intermediate rating and 4 pages were assigned the lowest rating. These results indicate that the BootCaT procedure is able to find relevant pages with high precision, and that the increase in number of retrieved pages in the second and third runs does not appear

to lower corpus quality too much.

6 Conclusion

Our experiments confirm that the BootCaT procedure, thanks to its modular and knowledge-poor nature, can be easily adapted to mine usable resources from typologically unrelated languages.

Future research will focus on the development of segmentation rules that avoid excessive over-segmentation and under-segmentation. We will also develop techniques to extract complex terms in a more systematic way. More generally, we would like to study how factors such as reference corpus and quality and number of iterations affect the results.

Given that the BootCaT tools and the other programs we used (ChaSen, UCS and `recode`) are freely available and open-source, we hope that interested researchers and language professionals will help to test, improve and extend the procedure.

References

- E. Banno, Y. Onno, Y. Sakane and C. Shinagawa. 1999. *Genki: An integrated course in elementary Japanese*. Tokyo: The Japan Times.
- M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. *LREC 2004*.
- S. Evert. 2004. *The Statistics of Word Cooccurrences: Bigrams and Collocations*. Ph.D. thesis, University of Stuttgart.
- A. Fujii and T. Ishikawa. Utilizing the World Wide Web as an encyclopedia: Extracting term descriptions from semi-structured texts. *ACL 2000*.
- R. Ghani, R. Jones, and D. Mladenic. 2001. Mining the web to create minority language corpora. *CIKM 2001*, 279–286.
- A. Kilgarriff and G. Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29:333–347.
- Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara. 2000. *Morphological analysis system ChaSen version 2.2.1 manual*. NIST Technical Report.
- P. Rayson and R. Garside. 2000. Comparing corpora using frequency profiling. *Proceedings of Workshop on Comparing Corpora of ACL 2000*, 1-6.