Case Frame Compilation from the Web using High-Performance Computing

Daisuke Kawahara, Sadao Kurohashi

The University of Tokyo
7-3-1 Hongo Bunkyo-ku, Tokyo, 113-8656, JAPAN
{kawahara, kuro}@kc.t.u-tokyo.ac.jp

Abstract

Case frames are important knowledge for a variety of NLP systems, especially when wide-coverage case frames are available. To acquire such large-scale case frames, it is necessary to automatically compile them from an enormous amount of corpus. In this paper, we consider the web as a corpus. We first build a huge text corpus from the web, and then construct case frames from the corpus. It is infeasible to do these processes by one CPU, and thus we employ a high-performance computing environment composed of 350 CPUs. The acquired corpus consists of 470M sentences, and the case frames compiled from them have 90,000 verb entries. The case frames contain most examples of usual use, and are ready to be applied to lots of NLP analyses and applications.

1. Introduction

As a first step in text understanding, it is necessary to capture various relations in the text, such as syntactic, predicate-argument and anaphoric relations. Revealing such relations requires wide-coverage knowledge like world knowledge that people have. One of such knowledge is "case frames", which describe what kinds of nouns are related to each verb. For example, let us show a case frame of the Japanese verb "*tsumu*" (load/accumulate):

***tsumu* (load)**	{*jyuugyoin* (employee), *driver*, $\cdots$}*ga*
	{*kuruma* (car), *truck*, $\cdots$}*ni*
	{*nimotsu* (baggage), *busshi* (supply)}*wo*,

where "*ga*", "*wo*" and "*ni*" are Japanese case-marking postpositions, which correspond to nominative, accusative and dative, respectively.

Such case frames can be utilized to improve not only fundamental analyses but also NLP applications such as information retrieval, automatic summarization and machine translation. To make practical use of case frames in these applications, wide-coverage case frames are required.

Thus far, typical case frames for important verbs have been elaborated by hand. It is, however, difficult or almost impossible to make wide-coverage case frames manually, because not only verbs but also nouns with copula demand case frames, and new words are coined every day. We consider automatically constructing case frames from a large corpus.

The wide spread of the Internet in recent years has made a large volume of texts available on the web. In this paper, we regard the web as a kind of text corpus, and construct case frames from it. The problem arising here is the computational cost of handling the vast web. It would take several years to compile case frames from hundreds of millions of web pages, even if the latest PC is used. To circumvent this problem, we employ a high-performance computing environment composed of 350 CPUs.

2. Building a Web Corpus

This section describes our method for building a Japanese text corpus that is used to construct case frames. We call this corpus "web corpus" hereafter.

2.1. Collecting Web Pages

We use web pages that were collected by a web crawler developed by (Takahashi et al., 2002). Since this crawler is oriented to collecting Japanese pages, it is suitable for our purpose. The number of the web pages collected is approximately 400M. They contain web pages written not only in Japanese but also in other languages.

2.2. Extracting Japanese Sentences from the Web Pages

The problem is how to collect Japanese sentences of good-quality from the web, where various languages and styles exist in web documents (HTML files). We process each web page using meta information of HTML and linguistic characteristics of Japanese as follows:

1. Extract candidates of Japanese web pages using encoding information

   (a) If a web page has "charset" information[1] and it corresponds to one of Japanese encodings (euc-jp, x-euc-jp, iso-2022-jp, shift_jis, windows-932, x-sjis, shift-jp, utf-8), this page is selected as a candidate of Japanese pages. Although utf-8 is an encoding of Unicode, a web page in utf-8 is also selected as a candidate.

   (b) A web page without "charset" information is processed by Encode::guess_encoding() function of Perl, which estimates the encoding of the web page using distinctive byte sequences of each encoding. A web page that is judged to be a Japanese encoding shown above is selected as a candidate.

[1] "charset" attribute in a meta tag of an HTML file.

2. Extract Japanese web pages using linguistic information

   The collected candidates of Japanese web pages may contain web pages that are not written in Japanese when the specified encoding or automatically judged encoding is incorrect or utf-8. To extract only Japanese web pages, page candidates are checked using the content ratio of the following Japanese postpositions:

   ga, *wo*, *ni*, *wa*, *no*, *de*

   We judge web pages whose content ratio of the postpositions exceeds 0.5% to be Japanese. As a result, 100M web pages were extracted as Japanese pages.

3. Split web pages into sentences

   The web pages are splitted into sentences using periods and HTML tags such as "br" and "p".

4. Extract Japanese sentences

   Even if a web page is judged to be Japanese, it may still contain sentences written in other languages. To obtain only Japanese sentences, we extract sentences that contain Japanese-specific characters such as HIRAGANA, KATAKANA and KANJI to a certain extent. We use the threshold of 60% out of the number of characters in each sentence. Finally, we discarded duplicate sentences, which might be extracted from a mirror site.

As a result, we acquired a corpus comprising 470M Japanese sentences. Table 1 shows a part of the web corpus.

2.3. Characteristics of the Acquired Web Corpus

We evaluated the quality of the acquired web corpus. We randomly selected 1,000 sentences from the web corpus, and examined whether each sentence is correct as Japanese or not. As a result, 995 sentences were correct. This result indicated that the quality of the acquired web corpus is really good.

We also examined some characteristics of the web corpus. First, we show how the number of unique words except unknown words change by increasing the corpus size (number of sentences). Figure 1 depicts it with one on a newspaper corpus. At the same corpus size, unique words of the web corpus exist more than those of the newspaper corpus. This indicates that the web corpus has wider coverage than the newspaper corpus. Figure 1 also shows the number of unique words occurring five times or more. Since the web corpus contains many low-frequency words, there is little difference between the web corpus and the newspaper corpus around 10 million sentences. However, according to the increase of corpus size, the number of unique words in the web corpus drastically increases.

Figure 2 shows the distribution of the sentence length, which is defined as the number of words in a sentence. This table indicates that the web corpus contains more short sentences than the newspaper corpus.

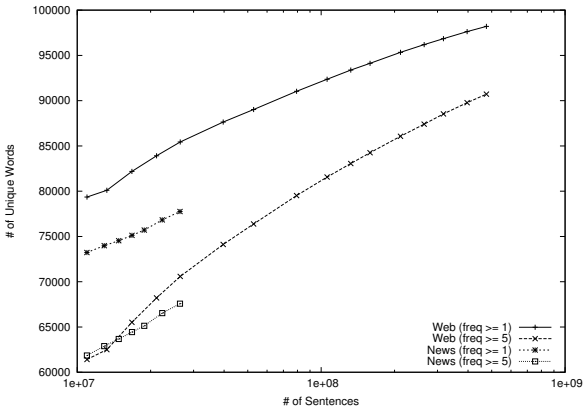

Figure 1: The relation between corpus size and the number of unique words.

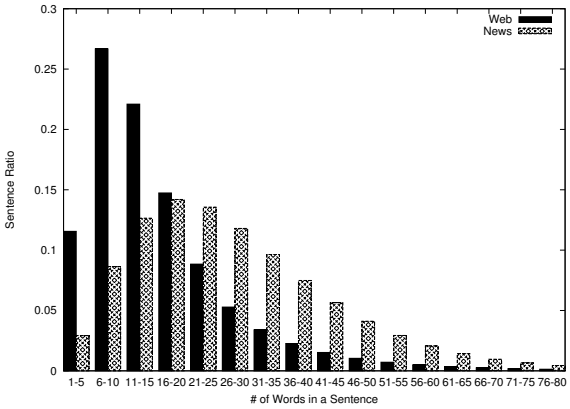

Figure 2: The distribution of sentence length.

3. The Method of Case Frame Construction

Case frames are constructed from modifier-head examples in automatic parses of the web corpus. The problems of automatic case frame construction are syntactic and semantic ambiguities. That is to say, the parsing results inevitably contain errors, and verb senses are intrinsically ambiguous. To cope with these problems, case frames are gradually constructed from reliable modifier-head examples (Kawahara and Kurohashi, 2002).

First, modifier-head examples that have no syntactic ambiguity are extracted, and they are disambiguated by a couple of a verb and its closest case component. Such couples are explicitly expressed on the surface of text, and can be considered to play an important role for constituting their sentence meanings. For instance, examples are distinguished not by verbs (e.g., "*tsumu*" (load/accumulate)), but by couples (e.g., "*nimotsu-wo tsumu*" (load baggage) and "*keiken-wo tsumu*" (accumulate experience)). Modifier-head examples are aggregated in this way, and yield basic case frames. Then, the basic case frames are clustered to merge similar case frames. For example, since "*nimotsu-wo tsumu*" (load baggage) and "*busshi-wo tsumu*" (load supply) are similar, and they are clustered. The similarity is measured using the thesaurus developed by NTT Communication Science Laboratories (NTT Communication Science Laboratories,

Table 1: Example sentences in the obtained web corpus.

もれなくプレゼント！ (Present it to you all!)
でも僕はＴシャツの上に長袖のシャツ。 (But, I wear a long-sleeved shirt on a T shirt.)
今回は某アイドルの高橋一也も参加したので客が若い。 (Since Kazuya Takahashi, who is an idol, joined this time, the audience was young.)
団体Ａが「まちづくり」をテーマにインターネット上で公開講座を開催しようとしている。 (The organization A is trying to hold an open class about "city planning" on the Internet.)
ｈｔａｃｃｅｓｓを置いたとたんそのディレクトリ以下で. (As soon as you put htaccess, under the directory.)
昨年の没後４００年祭を機に復元した井戸を紹介する木下さん (This is Mr. Kinoshita, who introduces a well restored last year marking fourth centennial of the death.)
恋は、真剣勝負。 (Love is a game played in earnest.)
ほめ言葉が多くって嬉しいですね。 (I'm glad to receive many compliments.)
いまだに言うでしょう。 (You still say that.)
「買いパラ」を見たと伝えれば、お買い上げ合計金額より５％引きいたします。 (If you say that you saw "*Kaipara*", we offer a 5% discount from all the bills.)
政治も危機的状況ですし、物資も不足しています。 (Politics is at a crisis, and commodities are scarce.)
思いやりのある優しい子に育ってネ。 (Grow up to be a considerate and kind person.)
毎月の費用もわずかです！ (Its monthly cost is very low!)

1997).

4. Case Frame Construction using HPC

To construct case frames by the method described above, we first apply morphological and syntactic analysis to the web corpus. We employ the Japanese morphological analyzer JUMAN and the syntactic analyzer KNP (Kurohashi and Nagao, 1998). Since KNP analyzes approximately 20 sentences per second, it impractically takes ten months to analyze the whole web corpus. To make this computation feasible, we employ a high-performance computing environment. The web corpus was divided into 10,000 pieces, and each of them were processed by a grid computing environment that consists of 350 CPUs. To submit these jobs to the grid, we used a grid shell GXP[2] (Kaneda et al., 2002). It took one day to finish the analyses.

Thereafter, we constructed case frames from the parsing results of the web corpus. Modifier-head examples were extracted from the parsing results, and were divided into each verb. This division made approximately 90,000 data, and they were used to construct case frames using the grid computing environment. This case frame construction took seven days.

5. Acquired Case Frames

Table 2 shows statistics of the acquired web case frames. For comparison, this table also shows statistics of the news case frames, which were built from 26M sentences in 26-year volumes of newspaper corpus. The statistics in Table 2 indicated that the acquired case frames are extremely larger than the case frames constructed from an existing corpus. We show some examples of the acquired web case frames in Table 3.

Table 2: Statistics of the acquired case frames of Web and News.

	Web	News
# of verbs	89243	18246
average # of case frames for a verb	34.3	17.5
average # of CS for a case frame	3.2	2.4
average # of examples for CS	72.9	29.8
average # of unique examples for CS	26.9	4.2

To investigate the coverage of the resultant case frames, we checked whether a predicate with its closest case component in a test sentence has a corresponding case frame. For test sentences, we used two sorts of text: 675 sentences from the web and 1,000 definition sentences from a Japanese dictionary for children. The result is shown in Table 4, where "exact" means that an exactly matched case frame exists, and "similar" means that a very similar case frame exists. The web case frames have wider coverage than the news case frames by more than 10%.

6. Previous Work

There has been some work extracting a text corpus from the web. Sekiguchi and Yamamoto built a Japanese web corpus (Sekiguchi and Yamamoto, 2004). They extracted Japanese sentences of good quality using HTML tags and

[2]http://www.logos.ic.i.u-tokyo.ac.jp/phoenix/

Table 3: Case frame examples (Examples are written only in English for space limitation. The number following each example means its frequency.).

	CS	examples
yaku (1) (broil)	*ga*	I:18, person:15, craftsman:10, ···
	wo	bread:2484, meat:1521, cake:1283, ···
	de	oven:1630, frying pan:1311, ···
yaku (2) (have difficulty)	*ga*	teacher:3, government:3, person:3, ···
	wo	fingers:2950
	ni	attack:18, action:15, son:15, ···
yaku (3) (burn)	*ga*	maker:1, distributor:1
	wo	data:178, file:107, copy:9, ···
	ni	R:1583, CD:664, CDR:3, ···
⋮	⋮	⋮
oyogu (1) (swim)	*ga*	dolphin:142, tutee:50, fish:28, ···
	wo	sea:1188, underwater:281, ···
	de	crawl:86, breaststroke:49, stroke:24, ···
⋮	⋮	⋮
migaku (1) (brush)	*ga*	I:4, man:4, person:4, ···
	wo	tooth:5959, molar:27, foretooth:12
	de	brush:38, salt:13, powder:12, ···
⋮	⋮	⋮
rokuga (1) (record)	*ga*	husband:4, sister:2, acquaintance:2, ···
	wo	program:1435, broadcast:521, ···
	ni	video:3753, disc:256, ···
⋮	⋮	⋮

Table 4: Examination of case frame existence in Web and News case frames.

	CF	exact	exact+similar
Web	WebCF	510/812 (0.628)	631/812 (0.777)
	NewsCF	372/812 (0.458)	526/812 (0.648)
Def	WebCF	269/449 (0.599)	375/449 (0.835)
	NewsCF	216/449 (0.481)	326/449 (0.726)

character types, and finally acquired a text corpus that consists of 220 MB. They evaluated it through some tasks like case frame construction, and claimed that the obtained corpus has larger coverage than newspaper corpora. The corpus seems to correspond to 10-year volume of newspaper corpora, and its size is only 1/50 of our web corpus.

On the other hand, there have been many studies that make use of word or phrase frequencies from search engines on the Internet, and apply them to a variety of NLP applications (Kilgarriff and Grefenstette, 2003). For example, Lapata and Keller utilized n-gram frequencies obtained from AltaVista in several NLP tasks such as word selection in machine translation, spelling correction and compound noun analysis (Keller and Lapata, 2003; Lapata and Keller, 2004). They indicated that a few tasks including word selection performed better than previous approaches based on existing text corpora, but many tasks did not achieve good performance. This is because such approaches cannot use linguistic information such as parts of speech and syntactic structure other than n-gram frequencies. Our web corpus can provide rich linguistic information by applying various NLP analyses, and can be employed in lots of NLP applications.

7. Conclusion

We have described a method for constructing wide-coverage case frames from the web. We first built a huge corpus from the web, and then constructed case frames from the corpus. The acquired corpus and case frames are extremely larger than previously built corpora and case frames. The resultant case frames contain most examples of usual use, and are ready to be applied to lots of NLP analyses and applications.

8. Acknowledgment

We would like to thank Prof. Kenjiro Taura for allowing us to use an enormous amount of web corpus and a grid computing environment.

9. References

Kenji Kaneda, Kenjiro Taura, and Akinori Yonezawa. 2002. Virtual private grid: A command shell for utilizing hundreds of machines efficiently. In *2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2002)*.

Daisuke Kawahara and Sadao Kurohashi. 2002. Fertilization of case frame dictionary for robust Japanese case analysis. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 425–431.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

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

Sadao Kurohashi and Makoto Nagao. 1998. Building a Japanese parsed corpus while improving the parsing system. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 719–724.

Mirella Lapata and Frank Keller. 2004. The web as a baseline: Evaluating the performance of unsupervised web-based models for a range of NLP tasks. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 121–128.

NTT Communication Science Laboratories. 1997. *Japanese Lexicon*. Iwanami Publishing.

Youichi Sekiguchi and Kazuhide Yamamoto. 2004. Improving quality of the web corpus. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*, pages 201–206.

Toshiyuki Takahashi, Hong Soonsang, Kenjiro Taura, and Akinori Yonezawa. 2002. World wide web crawler. In *Poster Proceedings of the 11th International World Wide Web Conference*.