PAPER Special Section on Info-Plosion

The Effect of Corpus Size on Case Frame Acquisition for Predicate-Argument Structure Analysis

Ryohei SASANO[†a)], **Daisuke KAWAHARA**[††], *Nonmembers*, *and* **Sadao KUROHASHI**[†], *Member*

SUMMARY This paper reports the effect of corpus size on case frame acquisition for predicate-argument structure analysis in Japanese. For this study, we collect a Japanese corpus consisting of up to 100 billion words, and construct case frames from corpora of six different sizes. Then, we apply these case frames to syntactic and case structure analysis, and zero anaphora resolution, in order to investigate the relationship between the corpus size for case frame acquisition and the performance of predicate-argument structure analysis. We obtained better analyses by using case frames constructed from larger corpora; the performance was not saturated even with a corpus size of 100 billion words.

key words: *corpus size, case frame, predicate-argument structure analysis*

1. Introduction

Very large corpora obtained from the Web have been successfully utilized for many natural language processing (NLP) applications, such as prepositional phrase (PP) attachment, other-anaphora resolution, spelling correction, confusable word set disambiguation and machine translation [1]–[5].

Most of the previous work utilized only the surface information of the corpora, such as *n*-grams, and co-occurrence counts. This may be because these studies did not require structured knowledge, and for such studies, the size of currently available corpora is considered to have been almost enough. For instance, while Brants et al. [5] reported that translation quality continued to improve with increasing corpus size for training language models at even size of 2 trillion tokens, the increase became small at the corpus size of larger than 30 billion tokens.

However, for more complex NLP tasks, such as case structure analysis and zero anaphora resolution, it is necessary to obtain structured knowledge, such as semantic case frames, which describe the cases each predicate has and the types of nouns that can fill a case slot. Note that case frames offer not only the knowledge of the relationships between a predicate and its particular case slot, but also the knowledge of the relationships among a predicate and its multiple case slots. To obtain such knowledge, very large corpora seem to be necessary; however it is still unknown how much corpora would be required to obtain good coverage. For example, Kawahara and Kurohashi proposed a method for constructing wide-coverage case frames from large corpora [6], and a model for syntactic and case structure analysis of Japanese that based upon case frames [7]. However, they did not demonstrate whether the coverage of case frames was wide enough for these tasks and how dependent the performance of the model was on the corpus size for case frame construction.

This paper aims to address these questions. For this purpose, we first collect a very large Japanese corpus consisting of about 100 billion words, or 1.6 billion unique sentences from the Web, and randomly select subsets of the corpus to obtain corpora of different sizes ranging from 1.6 million to 1.6 billion sentences. Then we construct case frames from each corpus by using Kawahara and Kurohashi's method [6], and apply them to existing models of syntactic and case structure analysis [7] and zero anaphora resolution [8], in order to investigate the relationship between the corpus size and the performance of these analyses.

Our main findings are as follows: better predicate-argument analysis is obtained by using case frames constructed from larger corpora; the performance is not saturated even with a corpus size of 100 billion words.

2. Related Work

Many NLP tasks have successfully utilized very large corpora, most of which were acquired from the Web [9]. Volk [1] proposed a method for resolving PP attachment ambiguities based upon Web data. Modjeska et al. [2] used the Web for resolving nominal anaphora. Lapata and Keller [3] investigated the performance of web-based models for a wide range of NLP tasks, such as MT candidate selection, article generation, and countability detection. Nakov and Hearst [10] solved relational similarity problems using the Web.

With respect to the effect of corpus size on NLP tasks, Banko and Brill [11] showed that for content sensitive spelling correction, increasing the training data size improved the accuracy. Atterer and Schütze [4] investigated the effect of corpus size in combining supervised and unsupervised learning for two types of attachment decision; they found that the combined system only improved the performance of the parser for small training sets. Brants et al. [5] varied the amount of language model training data from 13 million to 2 trillion tokens and applied these models to machine translation systems. They reported that translation

Manuscript received September 10, 2009.
Manuscript revised January 5, 2010.

[†]The authors are with the Graduate School of Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan.
[††]The author is with the National Institute of Information and Communications Technology, Kyoto-fu, 619–0289 Japan.
a) E-mail: sasano@i.kyoto-u.ac.jp

DOI: 10.1587/transinf.E93.D.1361

quality continued to improve with increasing corpus size for training language models at even size of 2 trillion tokens. Suzuki and Isozaki [12] provided evidence that the use of more unlabeled data in semi-supervised learning could improve the performance of NLP tasks, such as POS tagging, syntactic chunking, and named entity recognition.

There are several methods to extract useful information from very large corpora. Search engines, such as Google and Altavista, are often used to obtain Web counts (e.g. [13], [14]). However, search engines are not designed for NLP research and the reported hit counts are subject to uncontrolled variations and approximations. Therefore, several researchers have collected corpora from the Web by themselves. For English, Banko and Brill [15] collected a corpus with 1 billion words from variety of English texts. Liu and Curran [16] created a Web corpus for English that contained 10 billion words and showed that for content-sensitive spelling correction the Web corpus results were better than using a search engine. Halacsy et al. [17] created a corpus with 1 billion words for Hungarian from the Web by downloading 18 million pages. Others utilize publicly available corpus such as the North American News Corpus (NANC) and the Gigaword Corpus [18]. For instance, McClosky et al. [19] proposed a simple method of self-training a two phase parser-reranker system using NANC.

As for Japanese, Kawahara and Kurohashi [6] collected 23 million pages and created a corpus with approximately 20 billion words. Google released Japanese *n*-gram constructed from 20 billion Japanese sentences [20]. Several news wires are publicly available consisting of tens of million sentences. Kotonoha project is now constructing a balanced corpus of the present-day written Japanese consisting of 50 million words [21].

3. Construction of Case Frames

In this study, we construct case frames from raw corpora by using the method proposed by Kawahara and Kurohashi [6]. This section illustrates the methodology for constructing case frames.

3.1 Basic Method

After parsing a large corpus by a Japanese parser KNP[†], we construct case frames from predicate-argument examples in the resulting parses. The problems are syntactic and semantic ambiguities. In other words, the resulting parses inevitably contain errors and predicate senses are intrinsically ambiguous. To cope with these problems, we construct case frames from reliable predicate-argument examples.

First, we extract predicate-argument examples that had no syntactic ambiguity, and assemble them by coupling a predicate and its closest argument. That is, we assemble the examples not by predicates, such as *tsumu* (load/accumulate), but by couples, such as *nimotsu-wo tsumu* (load baggage) and *keiken-wo tsumu* (accumulate experience). Such couples are considered to play an important role for constituting sentence meanings. We call the assembled examples as basic case frames. Then, we cluster the basic case frames to merge similar case frames. For example, since *nimotsu-wo tsumu* (load baggage) and *busshi-wo tsumu* (load supplies) are similar, they are merged. The similarity is measured by using a Japanese thesaurus [22]. In order to remove inappropriate examples, we introduce a threshold α, and use only examples that appeared no less than α times in the corpora. In addition, in order to construct case frames within a practical time, we also introduce a threshold β, and eliminate such examples that are not the β most frequent examples, and are not merged with the β most frequent examples. Table 1 shows examples of constructed case frames.

3.2 Generalization of Examples

By using case frames automatically constructed from a large corpus, the data sparseness problem was alleviated, but not eliminated. For instance, there are thousands of named entities (NEs) that cannot be covered intrinsically. To deal with this problem, we generalize the examples of the case slots. Kawahara and Kurohashi also generalized examples but only for a few types. In this study, we generalize case slot examples based upon common noun categories and NE classes.

First, we generalize the examples based upon the categories that tagged by the Japanese morphological analyzer JUMAN[††]. In JUMAN, about 20 categories are defined and tagged to common nouns. For example, *ringo* (apple), *inu* (dog) and *byoin* (hospital) are tagged as FOOD, ANIMAL and FACILITY, respectively. For each category, we calculate the ratio of the categorized example among all case slot examples, and add it to the case slot (e.g. [CT:FOOD]:0.07).

We also generalize the examples based upon NE classes. We use a common standard NE definition for Japanese provided by the IREX [23]. We first recognize NEs in the source corpus by using an NE recognizer [24]; and then construct case frames from the NE-recognized corpus. Similar to the categories, for each NE class, we calculate the NE ratio among all the case slot examples, and add it to the case slot (e.g. [NE:PERSON]:0.12). The generalized examples are also included in Table 1.

4. Predicate-Argument Structure Analysis with Case Frames

In order to investigate the effect of corpus size on complex NLP tasks, we apply the constructed case frames to an integrated probabilistic model for Japanese syntactic and case structure analysis [7] and a probabilistic model for Japanese zero anaphora resolution [8]. In this section, we briefly describe these models.

[†]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp-e.html
[††]http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html

Table 1 Examples of constructed case frames.

	Case slot	Examples	Generalized examples with rate
tsumu (1) (load)	*ga* (nominative)	he, driver, friend, ···	[CT:PERSON]:0.45, [NE:PERSON]:0.08, ···
	wo (accusative)	baggage, luggage, hay, ···	[CT:ARTIFACT]:0.31, ···
	ni (dative)	car, truck, vessel, seat, ···	[CT:VEHICLE]:0.32, ···
tsumu (2) (accumulate)	*ga* (nominative)	player, children, party, ···	[CT:PERSON]:0.40, [NE:PERSON]:0.12, ···
	wo (accusative)	experience, knowledge, ···	[CT:ABSTRACT]:0.47, ···
⋮	⋮	⋮	
hanbai (1) (sell)	*ga* (nominative)	company, Microsoft, firm, ···	[NE:ORGANIZATION]:0.16, [CT:ORGANIZATION]:0.13, ···
	wo (accusative)	goods, product, ticket, ···	[CT:ARTIFACT]:0.40, [CT:FOOD]:0.07, ···
	ni (dative)	customer, company, user, ···	[CT:PERSON]:0.28, ···
	de (locative)	shop, bookstore, site ···	[CT:FACILITY]:0.40, [CT:LOCATION]:0.39, ···
⋮	⋮	⋮	

4.1 Model for Syntactic and Case Structure Analysis

Kawahara and Kurohashi [7] proposed an integrated probabilistic model for Japanese syntactic and case structure analysis based upon case frames. Case structure analysis recognizes predicate argument structures. Their model gives a probability to each possible syntactic structure T and case structure L of the input sentence S, and outputs the syntactic and case structure that have the highest probability. That is to say, the system selects the syntactic structure T_{best} and the case structure L_{best} that maximize the probability $P(T, L|S)$:

$$\begin{aligned}(T_{best}, L_{best}) &= \underset{(T,L)}{\operatorname{argmax}} P(T, L|S) \\ &= \underset{(T,L)}{\operatorname{argmax}} P(T, L, S) \end{aligned} \quad (1)$$

The last equation is derived because $P(S)$ is constant. $P(T, L, S)$ is defined as the product of a probability for generating a clause C_i as follows:

$$P(T, L, S) = \prod_{i=1,\ldots,n} P(C_i|b_{h_i}) \quad (2)$$

where n is the number of clauses in S, and b_{h_i} is C_i's modifying *bunsetsu*†. $P(C_i|b_{h_i})$ is approximately decomposed into the product of several generative probabilities such as $P(A(s_j) = 1|CF_l, s_j)$ and $P(n_j|CF_l, s_j, A(s_j) = 1)$, where the function $A(s_j)$ returns 1 if a case slot s_j is filled with an input argument; otherwise 0. $P(A(s_j) = 1|CF_l, s_j)$ denotes the probability that the case slot s_j is filled with an input argument, and is estimated from resultant case structure analysis of a large raw corpus. $P(n_j|CF_l, s_j, A(s_j) = 1)$ denotes the probability of generating a content part n_j from a filled case slot s_j in a case frame CF_l, and is calculated by using case frames. For details see [7].

4.2 Model for Zero Anaphora Resolution

Zero anaphora resolution is the integrated task of zero pronoun detection and zero pronoun resolution. In Japanese, since anaphors are often omitted, which are called *zero pronouns*, zero anaphora resolution is one of the most important techniques for discourse analysis.

We proposed a probabilistic model for Japanese zero anaphora resolution based upon case frames [8]. This model first resolves coreference and identifies discourse entities; then gives a probability to each possible case frame CF and case assignment CA when target predicate v, input arguments IA and existing discourse entities ENT are given, and outputs the case frame and case assignment that have the highest probability. That is to say, this model selects the case frame CF_{best} and the case assignment CA_{best} that maximize the probability $P(CF, CA|v, IA, ENT)$:

$$\begin{aligned}&(CF_{best}, CA_{best}) \\ &\quad = \underset{(CF,CA)}{\operatorname{argmax}} P(CF, CA|v, IA, ENT)\end{aligned} \quad (3)$$

$P(CF, CA|v, IA, ENT)$ is approximately decomposed into the product of several probabilities. Case frames are used for calculating $P(n_j|CF_l, s_j, A(s_j) = 1)$, the probability of generating a content part n_j from a case slot s_j in a case frame CF_l, and $P(n_j|CF_l, s_j, A'(s_j) = 1)$, the probability of generating a content part n_j of a zero pronoun, where the function $A'(s_j)$ returns 1 if a case slot s_j is filled with an antecedent of a zero pronoun; otherwise 0. $P(n_j|CF_l, s_j, A'(s_j) = 1)$ is similar to $P(n_j|CF_l, s_j, A(s_j) = 1)$ and estimated from the frequencies of case slot examples in case frames. However, while $A'(s_j) = 1$ means s_j is not filled with an overt argument but filled with an antecedent of zero pronoun, case frames are constructed from overt predicate argument pairs. Therefore, the content part n_j is often not included in the case slot examples. To cope with this problem, this model also utilizes generalized examples to estimate $P(n_j|CF_l, s_j, A(s_j) = 1)$. For details see [8].

5. Experiments

5.1 Construction of Case Frames

In order to investigate the effect of corpus size, we constructed case frames from corpora of different sizes. We first collected Japanese sentences from the Web using the method proposed by Kawahara and Kurohashi [6]. We acquired approximately 6 billion Japanese sentences consisting of approximately 100 billion words from 100 million

†In Japanese, *bunsetsu* is a basic unit of dependency, consisting of one or more content words and the following zero or more function words. It corresponds to a base phrase in English.

Japanese web pages. After discarding duplicate sentences, which would have been extracted from mirror sites, we acquired a corpus comprising of 1.6 billion (1.6G) unique Japanese sentences consisting of approximately 25 billion words. The average number of characters and words in each sentence was 28.3, 15.6, respectively. Then we randomly selected subsets of the corpus for five different sizes; 1.6M, 6.3M, 25M, 100M, and 400M sentences to obtain corpora of different sizes.

We constructed case frames from each corpus. We employed JUMAN and KNP to parse each corpus. We changed the threshold α introduced in Sect. 3.1 depending upon the size of the corpus as shown in Table 2, and fixed the threshold β at 1,000. Completing the case frame construction took about two weeks using 600 CPUs. Table 3 shows the statistics for the constructed case frames. The number of predicates, the average number of examples and unique examples for a case slot, and whole file size were confirmed to be heavily dependent upon the corpus size. However, the average number of case frames for a predicate and case slots for a case frame did not.

5.2 Coverage of Constructed Case Frames

5.2.1 Setting

In order to investigate the coverage of the resultant case frames, we used a syntactic relation, case structure, and anaphoric relation annotated corpus consisting of 186 web documents (979 sentences). This corpus was manually annotated using the same criteria as Kawahara et al. [25]. There were 2,390 annotated relations between predicates and their direct (not omitted) arguments and 837 zero anaphoric relations in the corpus.

We used two evaluation metrics depending upon whether the target argument was omitted or not. For overt argument, we judged the target argument was covered by case frames if the argument itself was included in the examples for one of the corresponding case slots of the case frame. For omitted argument, we checked not only the target argument itself but also all mentions that refer to the same entity as the argument.

Table 2 Corpus sizes and thresholds.

Corpus size for case frame construction (sentences)	1.6M	6.3M	25M	100M	400M	1.6G
Threshold α introduced in Sect. 3.1	2	3	4	5	7	10
Corpus size to estimate generative probability (sentences)	1.6M	3.2M	6.3M	13M	25M	50M

5.2.2 Coverage of Case Frames

Figure 1 shows the coverage of case frames for the overt argument, which would have tight relations with case structure analysis. The lower line shows the coverage without considering generalized examples, the middle line shows the coverage considering generalized NE examples, and the upper line shows the coverage considering all generalized exam-

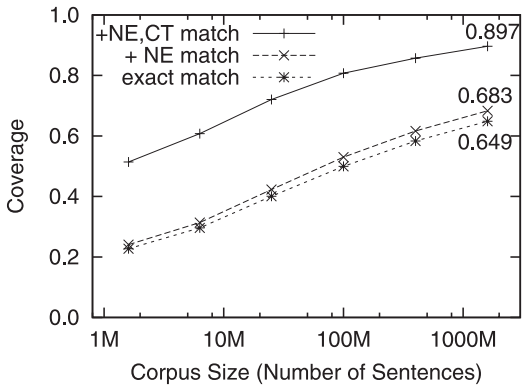


Fig. 1 Coverage of CF (overt argument).

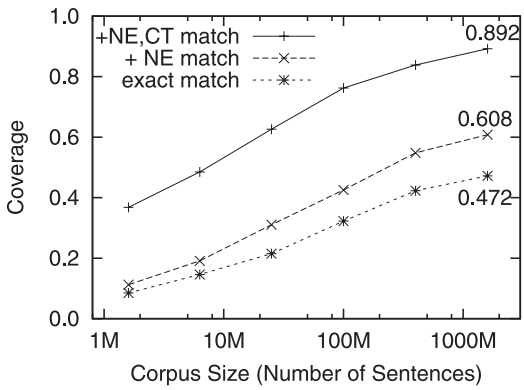


Fig. 2 Coverage of CF (omitted argument).

Table 3 Statistics of the constructed case frames.

Corpus size (sentences)		1.6M	6.3M	25M	100M	400M	1.6G
# of predicate		2460	6134	13532	27226	42739	65679
(type)	verb	2039	4895	10183	19191	28523	41732
	adjective	154	326	617	1120	1641	2318
	noun with copula	267	913	2732	6915	12575	21629
average # of case frames for a predicate		15.9	12.2	13.3	16.1	20.5	25.3
average # of case slots for a case frame		2.95	3.44	3.88	4.21	4.69	5.08
average # of examples for a case slot		4.89	10.2	19.5	34.0	67.2	137.6
average # of unique examples for a case slot		1.19	1.85	3.06	4.42	6.81	9.64
average # of generalized examples for a case slot		0.14	0.24	0.37	0.49	0.67	0.84
File size (byte)		8.9M	20M	56M	147M	369M	928M

ples.

Figure 2 shows the coverage of case frames for the omitted argument, which would have tight relations with zero anaphora resolution. The upper line shows the coverage considering all generalized examples, which is considered to be the upper bound of performance for the zero anaphora resolution system described in Sect. 4.2.

Both figures show that the coverage was improved by using larger corpora and there was no saturation even when the corpus of 1.6 billion sentences was used. When the largest corpus and all generalized examples were used, the case frames achieved a coverage of almost 90% for both the overt and omitted argument.

These figures also suggest that generalization of NEs is less effective than that of categories. However, while about half of the predicate-argument pairs that are not covered by case frames or only covered with considering generalized categories appear one or more times in the source corpus but are filtered by the thresholds α or β, only about one out of four predicate-argument pairs that are only covered with considering generalized NEs appear in the source corpus. Thus, we can say that most examples of NEs cannot be collected and the generalization of NEs is intrinsically important. We show some examples of predicate-argument pairs that are not covered by case frames in Table 4.

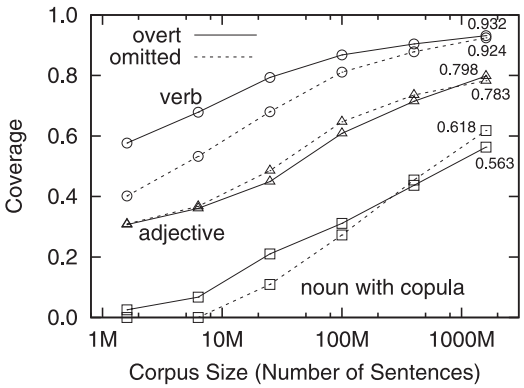


Fig. 3 Coverage of CF for each predicate type considering all generalized examples.

Comparing Fig. 2 with Fig. 1, we found two characteristics. First, the lower and middle lines of Fig. 2 were located lower than the corresponding lines in Fig. 1. This would reflect that some frequently omitted arguments are not described in the case frames because the case frames were constructed from only overt predicate argument pairs. Secondly, the effect of generalized NE examples was more evident for the omitted argument, which would reflect the important role of NEs in zero anaphora resolution.

Figure 3 shows the coverage of case frames for each predicate type, which was calculated with considering all generalized examples. The case frames for verbs achieved a coverage of about 93%. For adjective, the coverage was about 78%. The main cause of the lower coverage was that the predicate argument relations concerning adjectives that were used in restrictive manner, such as *yûgana hadaai* (elegant texture) (cf. (8) in Table 4), were not used for case frame construction, although such relations were also the target of the coverage evaluation. For noun with copula, the coverage was only about 60%. However, most predicate argument relations concerning nouns with copula were easily recognized from syntactic preference (cf. (9) and (10) in Table 4), and thus the low coverage would not quite affect the performance of predicate-argument structure analysis.

5.3 Syntactic and Case Structure Analysis

5.3.1 Accuracy of Syntactic Analysis

We investigated the effect of corpus size for syntactic analy-

Table 4 Examples of predicate-argument pairs that are not covered by case frames, including some examples that are only covered with considering generalized examples.

	POS of the predicate	Argument type	Example (argument)	(predicate)	Others
(1)	verb	overt	*nechizun-ga* netizen	*tairitsu-suru* conflict	No example.
(2)	verb	overt	*hokenjo-ga* health centre	*youbou-suru* request	4 examples that are filtered by the threshold α.
(3)	verb	overt	*shikisha-wo* conductor	*imêji-suru* image	4 examples that are filtered by the threshold α. Covered with considering generalized categories.
(4)	verb	omitted	*watashi-ni* me	*toiawasu* inquire	No example.
(5)	verb	omitted	*heisha-ga* our company	*negau* hope	3 examples that are filtered by the threshold α.
(6)	verb	omitted	*Kinjô-ga* Kinjô	*tsutomeru* work	No example. Covered with considering generalized NEs.
(7)	adjective	overt	*shushi-ga* seed	*kuroi* black	No example. Covered with considering generalized categories.
(8)	adjective	omitted	*hadaai-ga* texture	*yûga* elegant	No example.
(9)	noun with copula	overt	*Irifunesô-ga* Irifunesô	*yado-da* inn	No example. Covered with considering generalized NEs.
(10)	noun with copula	omitted	*hatsuden-ga* power generation	*shikumi-da* mechanism	No example.

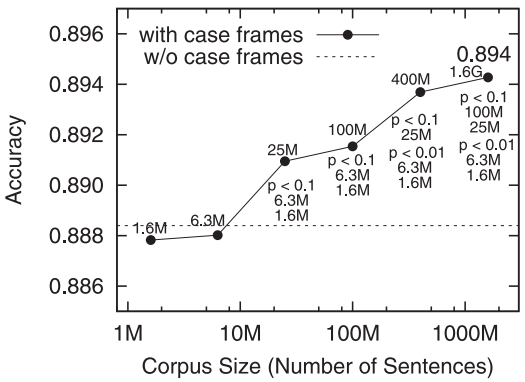


Fig. 4 Accuracy of syntactic analysis. (McNemar's test results are also shown under each data point. For instance, the difference of 1.6G from 400M is not significant at even the 90% level ($p = 0.1$), the differences from 100M and 25M are significant at the 90% level, but not significant at the 99% level ($p = 0.01$), the differences from 6.3M and 1.6M are significant at even 99% level.)

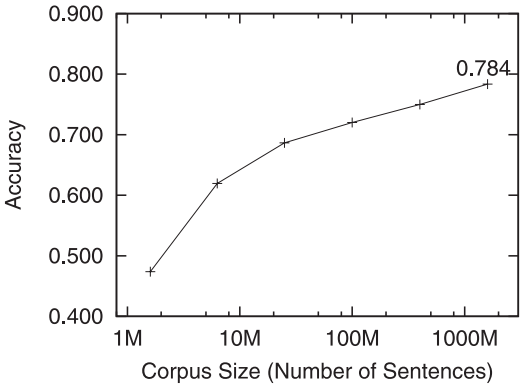


Fig. 5 Accuracy of case structure analysis.

Table 5 Corpus sizes for case frame construction and time for syntactic and case structure analysis.

Corpus size	1.6M	6.3M	25M	100M	400M	1.6G
Time (sec.)	850	1244	1833	2696	3783	5553

sis described in Sect. 4.1. We used hand-annotated 759 web sentences for evaluation, which was used by Kawahara and Kurohashi [26]. The unlexical parameters were calculated from the Kyoto Text Corpus [27], which consists of 40K Japanese newspaper sentences, and is syntactically annotated in dependency formalism. We evaluated the resultant syntactic structures with regard to dependency accuracy, the proportion of correct dependencies out of all dependencies[†].

Figure 4 shows the accuracy of syntactic structures. We conducted these experiments with case frames constructed from corpora of different sizes. We also changed the corpus size to estimate generative probability of a case slot in Sect. 4.1 depending upon the size of the corpus for case frame construction as shown in Table 2. Figure 4 also includes McNemar's test results.

In Fig. 4, 'w/o case frames' shows the accuracy of the rule-based syntactic parser KNP that does not use case frames. Since the model described in Sect. 4.1 assumes the existence of reasonable case frames, when we used case frames constructed from very small corpus, such as 1.6M and 6.3M sentences, the accuracy was lower than that of the rule-based syntactic parser.

We confirmed that better performance was obtained by using case frames constructed from larger corpora, and the accuracy of 0.894[††] was achieved by using the case frames constructed from 1.6G sentences. However the effect of the corpus size was limited. This is because there are various causes of dependency error and the case frame sparseness problem is not serious for syntactic analysis.

We considered that generalized examples can benefit for the accuracy of syntactic analysis, and tried several models that utilize these examples. However, we cannot confirm any improvement.

5.3.2 Accuracy of Case Structure Analysis

We conducted case structure analysis on 215 web sentences in order to investigate the effect of corpus size for case structure analysis. The case markers of topic marking phrases and clausal modifiers were evaluated by comparing them with the gold standard in the corpus. Figure 5 shows the experimental results. We confirmed that the accuracy of case structure analysis strongly depends on corpus size for case frame construction.

5.3.3 Analysis Speed

Table 5 shows the time for analyzing syntactic and case structure of 759 web sentences. Although the time for analysis became longer by using case frames constructed from a larger corpus, the growth rate was smaller than the growth rate of the size for case frames described in Table 3.

Since there is enough increase in accuracy of case structure analysis, we can say that case frames constructed larger corpora are desirable for case structure analysis.

5.4 Zero Anaphora Resolution

5.4.1 Accuracy of Zero Anaphora Resolution

We used an anaphoric relation annotated corpus consisting of 186 web documents (979 sentences) to evaluate zero anaphora resolution. We used first 51 documents for test and used the other 135 documents for calculating several probabilities. In the 51 test documents, 233 zero anaphora relations were annotated between one of the mentions of the antecedent and corresponding predicate that had zero pronoun.

In order to concentrate on evaluation for zero anaphora

[†]Note that Kawahara and Kurohashi [26] exclude the dependency between the last two *bunsetsu*, since Japanese is head-final and thus the second last *bunsetsu* unambiguously depends on the last *bunsetsu*.

[††]It corresponds to 0.877 in Kawahara and Kurohashi's [26] evaluation metrics.

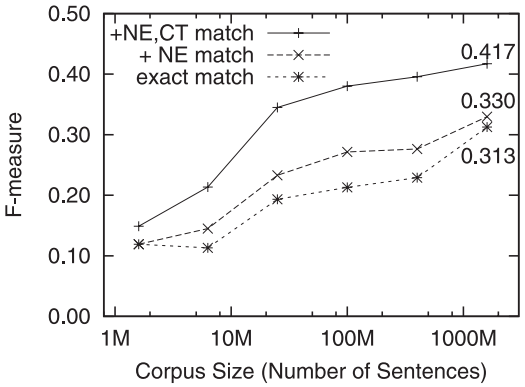


Fig. 6 F-measure of zero anaphora resolution.

Table 6 Corpus sizes for case frame construction and time for zero anaphora resolution.

Corpus size	1.6M	6.3M	25M	100M	400M	1.6G
Time (sec.)	538	545	835	1040	1646	2219

resolution, we used the correct morphemes, named entities, syntactic structures and coreference relations that were manually annotated. Since correct coreference relations were given, the number of created entities was the same between the gold standard and the system output because zero anaphora resolution did not create new entities.

The experimental results are shown in Fig. 6, in which F-measure was calculated by:

$$R = \frac{\text{\# of correctly recognized zero anaphora}}{\text{\# of zero anaphora annotated in corpus}},$$

$$P = \frac{\text{\# of correctly recognized zero anaphora}}{\text{\# of system outputted zero anaphora}},$$

$$F = \frac{2}{1/R + 1/P}.$$

The upper line shows the performance using all generalized examples, the middle line shows the performance using only generalized NEs, and the lower line shows the performance without using any generalized examples. While generalized categories much improved the F-measure, generalized NEs contributed little. This tendency is similar to that of coverage of case frames for omitted argument shown in Fig. 2. Unlike syntactic and case structure analysis, the performance for the zero anaphora resolution is quite low when using case frames constructed from small corpora, and we can say case frames constructed from larger corpora are essential for zero anaphora resolution.

5.4.2 Analysis Speed

Table 6 shows the time for resolving zero anaphora in 51 web documents consisting of 278 sentences. The time for analysis became longer by using case frames constructed from larger corpora, which tendency is similar to the growth of the time for analyzing syntactic and case structure.

5.5 Discussion

Experimental results of both case structure analysis and zero anaphora resolution show the effectiveness of a larger corpus in case frame acquisition for Japanese predicate-argument structure analysis. Up to the corpus size of 1.6 billion sentences, or 100 billion words, these experimental results still show a steady increase in performance. That is, we can say that the corpus size of 1.6 billion sentences is not enough to obtain case frames of sufficient coverage.

These results suggest that increasing corpus size is more essential for acquiring structured knowledge than for acquiring unstructured statistics of a corpus, such as *n*-grams, and co-occurrence counts; and for complex NLP tasks such as case structure analysis and zero anaphora resolution, the currently available corpus size is not sufficient.

Therefore, to construct more wide-coverage case frames by using a larger corpus and reveal how much corpora would be required to obtain sufficient coverage is considered as future work.

6. Conclusion

This paper has reported the effect of corpus size on case frame acquisition for syntactic and case structure analysis, and zero anaphora resolution in Japanese. We constructed case frames from corpora of six different sizes ranging from 1.6 million to 1.6 billion sentences; and then applied these case frames to Japanese syntactic and case structure analysis, and zero anaphora resolution. Experimental results showed better analyses were obtained using case frames constructed from larger corpora, and the performance showed no saturation even when the corpus size was 1.6 billion sentences.

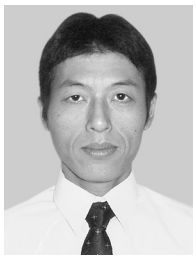
The findings suggest that increasing corpus size is more essential for acquiring structured knowledge than for acquiring surface statistics of a corpus; and for complex NLP tasks the currently available corpus size is not sufficient.

References

[1] M. Volk, "Exploiting the WWW as a corpus to resolve PP attachment ambiguities," Proc. Corpus Linguistics, pp.601–606, 2001.
[2] N.N. Modjeska, K. Markert, and M. Nissim, "Using the web in machine learning for other-anaphora resolution," Proc. EMNLP-2003, pp.176–183, 2003.
[3] M. Lapata and F. Keller, "Web-based models for natural language processing," ACM Trans. Speech and Language Processing, vol.2, issue 1, pp.1–31, 2005.
[4] M. Atterer and H. Schütze, "The effect of corpus size in combining supervised and unsupervised training for disambiguation," Proc. COLING-ACL'06, pp.25–32, 2006.
[5] T. Brants, A.C. Popat, P. Xu, F.J. Och, and J. Dean, "Large language models in machine translation," Proc. EMNLP-CoNLL'07, pp.858–867, 2007.
[6] D. Kawahara and S. Kurohashi, "Case frame compilation from the web using high-performance computing," Proc. LREC'06, pp.1344–1347, 2006.

[7] D. Kawahara and S. Kurohashi, "A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis," Proc. HLT-NAACL'06, pp.176–183, 2006.
[8] R. Sasano, D. Kawahara, and S. Kurohashi, "A fully-lexicalized probabilistic model for Japanese zero anaphora resolution," Proc. COLING'08, pp.769–776, 2008.
[9] A. Kilgarriff and G. Grefenstette, "Introduction to the special issue on the web as corpus," Computational Linguistic, vol.29, no.3, pp.333–347, 2003.
[10] P. Nakov and M.A. Hearst, "Solving relational similarity problems using the web as a corpus," Proc. ACL-HLT'08, pp.452–460, 2008.
[11] M. Banko and E. Brill, "Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing," Proc. HLT'01, 2001.
[12] J. Suzuki and H. Isozaki, "Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data," Proc. ACL-HLT'08, pp.665–673, 2008.
[13] P. Nakov and M. Hearst, "A study of using search engine page hits as a proxy for *n*-gram frequencies," Proc. RANLP'05, 2005.
[14] A. Gledson and J. Keane, "Using web-search results to measure word-group similarity," Proc. COLING'08, pp.281–288, 2008.
[15] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," Proc. ACL'01, pp.26–33, 2001.
[16] V. Liu and J.R. Curran, "Web text corpus for natural language processing," Proc. EACL'06, pp.233–240, 2006.
[17] P. Halacsy, A. Kornai, L. Nemeth, A. Rung, I. Szakadat, and V. Tron, "Creating open language resources for Hungarian," Proc. LREC'04, pp.203–210, 2004.
[18] D. Graff, "English gigaword," Technical Report LDC2003T05, Linguistic Data Consortium, Philadelphia, PA, USA, 2003.
[19] D. McClosky, E. Charniak, and M. Johnson, "Effective self-training for parsing," Proc. HLT-NAACL'06, pp.152–159, 2006.
[20] T. Kudo and H. Kazawa, Web Japanese N-gram version 1, Gengo Shigen Kyokai, 2007.
[21] K. Maekawa, "Kotonoha, the corpus development project of the National Institute for Japanese Language," Proc. 13th NIJL International Symposium, pp.55–62, 2006.
[22] The National Language Institute for Japanese Language, Bunruigoihyo, Dainippon Tosho, 2004.
[23] IREX Committee, ed., Proc. IREX Workshop, 1999.
[24] R. Sasano and S. Kurohashi, "Japanese named entity recognition using structural natural language processing," Proc. IJCNLP'08, pp.607–612, 2008.
[25] D. Kawahara, R. Sasano, and S. Kurohashi, "Toward text understanding: Integrating relevance-tagged corpora and automatically constructed case frames," Proc. LREC'04, pp.1833–1836, 2004.
[26] D. Kawahara and S. Kurohashi, "Probabilistic coordination disambiguation in a fully-lexicalized Japanese parser," Proc. EMNLP-CoNLL'07, pp.306–314, 2007.
[27] S. Kurohashi and M. Nagao, "Building a Japanese parsed corpus while improving the parsing system," Proc. LREC'98, pp.719–724, 1998.

Ryohei Sasano received the B.S., M.S., and Ph.D. in Information Science and Technology from the University of Tokyo in 2004, 2006, and 2009, respectively. He is currently a researcher of the Graduate School of Informatics at Kyoto University. His research interest is language processing, in particular anaphora resolution.



Daisuke Kawahara received his B.S. and M.S. in Electronic Science and Engineering from Kyoto University in 1997 and 1999, respectively. He obtained his Ph.D. in Informatics from Kyoto University in 2005. He is currently a researcher of National Institute of Information and Communications Technology, Japan. His research interests center on natural language processing, in particular knowledge acquisition and text understanding.

Sadao Kurohashi received the B.S., M.S., and Ph.D. in Electrical Engineering from Kyoto University in 1989, 1991 and 1994, respectively. He has been a visiting researcher of IRCS, University of Pennsylvania in 1994. He is currently a professor of the Graduate School of Informatics at Kyoto University. His research interests include natural language processing, knowledge acquisition, and information retrieval.