Annotating Syntactic Information on 5.5 Billion Word Corpus of Japanese Blogs

Michal Ptaszynski †　　Rafal Rzepka ‡　　Kenji Araki ‡　　Yoshio Momouchi §

† High-Tech Research Center, Hokkai-Gakuen University
ptaszynski@hgu.jp

‡ Graduate School of Information Science and Technology, Hokkaido University
{kabura,araki}@media.eng.hokudai.ac.jp

§ Department of Electronics and Information Engineering,
Faculty of Engineering, Hokkai-Gakuen University
momouchi@eli.hokkai-s-u.ac.jp

Abstract

In this paper we report on annotating syntactic information on YACIS, a 5.5 billion word corpus of Japanese blogs. The annotated information includes such features as tokenization, part-of-speech tagging, lemmatization, dependency parsing or detection of sentence boundaries. We present the statistics of those annotations and compare them with other corpora.

1 Introduction

The importance of text corpora is widely recognized in the field of Natural Language Processing (NLP), and numerous corpora have been compiled so far for different languages. However, comparing to major world languages, like English, there are few large corpora available for the Japanese language [1]. Moreover, grand majority of them is solely based on classical literature [2], newspapers [3], or legal documents. These usually do not contain casual language, the kind of language that appears most often in everyday use. Recently blogs have become recognized as a rich source of casual language. Blogs are open access Internet diaries in which people expressively describe their experiences or opinions. Thus the contents of blog posts have come into the focus of NLP [4, 5, 6]. Therefore creating a large blog-based corpus and annotating it with linguistic information could become a solution to overcome the problem of lack of casual language corpora. Although there exist several somewhat large web-based corpora (containing several million words), such as JpWaC [1], jBlogs [7] or KWIC on WEB [8], access to them is usually allowed only from the Web interface, which hinders additional annotations (dependency structure, named entity recognition, etc.). Therefore there was a need for a large-scale blog corpus annotated with linguistic information capable to be queried locally (e.g., when looking for word or sentence patterns), instead of querying the Internet through search engines. Maciejewski et al. [9] developed YACIS, a sufficiently large blog corpus. Unfortunately, the corpus was not annotated. We decided to annotate YACIS with all available linguistic information. In the annotation process we included part-of-speech (POS) tagging, dependency structure (DS) analysis and named entity recognition (NER).

The outline of this paper is as follows. In seciton 2 we briefly describe YACIS. Section 3 describes the tools we used to perform the annotations. In section 4 we present some of the statistics of the annotations and compare them to other corpora. Finally, we conclude the paper, and describe further work that needs to be performed on YACIS.

2 YACIS Corpus Description

YACIS or **Y**et **A**nother **C**orpus of **I**nternet **S**entences was collected automatically by Maciejewski et al. [9] from the pages of Ameba blog service. It contains 5.6 billion words within 350 million sentences. The compilation process was performed within 3 weeks between 3rd and 24th of December 2009. Maciejewski et al. extracted only pages containing Japanese posts (pages with legal disclaimers or written in languages other than Japanese were omitted). In the initial phase they provided their crawler, optimized to crawl only Ameba blog service, with 1000 links taken from Google (response to one simple query: 'site:ameblo.jp'). They saved all pages to disk as raw HTML files (each page in a separate file) and afterward extracted all the posts and comments and divided them into sentences. The original structure (blog post and comments) is preserved, thanks to which semantic relations between posts and comments are retained. Each blog page was transformed into an independent XML block between <doc></doc> tags. Opening tag of the <doc> block contains three parameters: URL, TIME

```
<doc url="http://ameblo.jp/blog-name/entry-000001.html"
time="2009-12-05 21:11:46" id="2000001">
   <post>
        <s>今日から十月です。</s>
        [Its October from today.]
        <s>なんか、九月はいつもよりアッという間に過ぎたような気がするなぁ。</s>
        [I have a strange feeling September passed faster than usual.]
        ...
   </post>
   <comments>
        <cmt>
                <s>色々と忙しいですね～！</s>
                [Oh, you've been busy, weren't you?]
                ...
        </cmt>
        <cmt>
                <s>お疲れ艹です(^o^)</s>
                [Well done! Cheers for good work (^o^)]
                ...
        </cmt>
   </comments>
</doc>
```

Figure 1: The example of YACIS XML structure.

Table 1: General Statistics of YACIS.

# of web pages	12,938,606
# of unique bloggers	60,658
average # of pages/blogger	213.3
# of pages with comments	6,421,577
# of comments	50,560,024
average # of comment/page	7.873
# of words/tokens	5,600,597,095
# of sentences	354,288,529
# of words per sentence (average)	15
# of characters per sentence (average)	77

and ID which specify the exact address from which the page was downloaded, download time and unique page index, respectively. The <doc> block contains two other tags: <post> and <comments>. The former contains all the sentences from the main post with each sentence included between <s></s> tags. The latter contains all comments written under the post, each placed between <cmt></cmt> tags split into sentences. An example of the XML structure is shown in figure 1. The corpus is stored in 129 text files containing 100,000 <doc> units each, and is encoded using UTF-8 encoding. The size of each file varies and is between 200 and 320 megabytes. The size of raw corpus (pure text corpus without any additional tags) is 27.1 gigabytes. Other primary statistics of the corpus are represented in table 1.

3 Syntactic Information Annotation Tools

The corpus in the form described in section 2 was further annotated with several kinds if information. We performed tokenization, POS tagging and lemmatization with MeCab, and dependency parsing and named entity recognition with Cabocha. Both tools are described in detail below.

MeCab [10] is a standard morphological analyzer and POS tagger for Japanese. It is trained using a large corpus on a Conditional Random Fields (CRF) discriminative model and uses a bigram Markov model for analysis. Except MeCab there are several POS taggers for Japanese,

Table 2: Named entity tags included in IREX.

<opening tag>...</closing tag>	**explanation**
<ORGANIZATION>... ...</ORGANIZATION>	organization or company name including abbreviations (e.g., Toyota, or Nissan);
<LOCATION>...</LOCATION>	mane of a place (city, country, etc.);
<PERSON>...</PERSON>	name, nickname, or status of a person (e.g., “me”, “grandson”, etc.);
<ARTIFACT>...</ARTIFACT>	name of a well recognized product or object (e.g., Van Houtens Cocoa, etc.);
<PERCENT>...</PERCENT>	percentage or ratio (90%, 0.9);
<MONEY>...</MONEY>	currencies (1000 $, 100 ¥);
<DATE>...</DATE>	dates and its paraphrased extensions (e.g., “4th July”, or “next season”, etc.)
<TIME>...</TIME>	hours, minutes, seconds, etc.

such as Juman[1] or ChaSen[2]. ChaSen and MeCab have many similarities in their structure. Both share the same corpus base for training and use the same default dictionary (ipadic[3] based on a modified IPA Part of Speech Tagset developed by the Information-Technology Promotion Agency of Japan (IPA)). However, ChaSen was trained on a Hidden Markov Model (generative model), a full probabilistic model in which first all variables are generated. Therefore it is about 3-4 times slower than MeCab, which is based on a discriminative model, in which focus is only on the target variables conditional to the observed variables. Juman on the other hand was developed separately from MeCab on different resources. It uses a set of hand-crafted rules and a dictionary (jumandic) created on the basis of Kyoto Corpus developed by a Kurohashi&Kawahara Laboratory[4] at Kyoto University. Both MeCab and Juman are considerably fast, which is a very important feature when processing a large-scale corpus such as YACIS. However, there were several reasons to choose the former. MeCab is considered slightly faster when processing large data and uses less memory. It is also more accurate since it allows partial analysis (a way of flexible setting of word boundaries in non-spaced languages, like Japanese). Finally, MeCab is flexible when using other dictionaries. Therefore to annotate YACIS we were able to use MeCab with the two different types of dictionaries mentioned above (ipadic and jumandic). This allowed us to develop POS tagging for YACIS with the two most favored standards in morphological analysis of Japanese today. An example of MeCab output using the ipadic dictionary is represented in figure 2.

Cabocha [11] is a Japanese dependency parser based on Support Vector Machines. It was developed by MeCab developers and is considered to be the most accurate statistical Japanese dependency parser. Its discriminative feature is using Cascaded Chunking Model, which makes the analysis efficient for the Japanese language. The Cascaded Chunking Model parses a sentence deterministically focusing on whether a sentence segment modifies

[1] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN
[2] http://chasen.naist.jp/hiki/ChaSen/
[3] http://sourceforge.jp/projects/ipadic/
[4] http://nlp.ist.i.kyoto-u.ac.jp/index.php

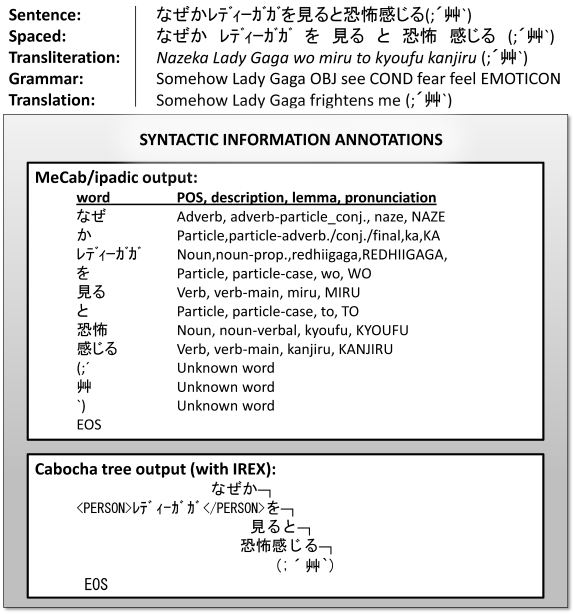

Figure 2: Output examples for MeCab and Cabocha.

a segment on its right hand side [11]. As an option, Cabocha uses IREX[5] (Information Retrieval and Extraction Exercise) standard for Named Entity Recognition (NER). We applied this option in the annotation process as well. Table 2 represents all types of tags included in IREX. An example of Cabocha output is represented in figure 2.

4 Syntactic Information Statistics

In this section we present all relevant statistics concerning syntactic information annotated on YACIS corpus. Where it was possible we also compared YACIS to other corpora. All basic information concerning YACIS is represented in table 1. Information on the distribution of parts of speech is represented in table 3. We compared the two dictionaries used in the annotation (ipadic and jumandic) with other Japanese corpora (jBlogs, and JENAAD newspaper corpus) and in addition, partially to British and Italian Web corpus (ukWaC and itWaC, respectively). The results of analysis are explained below.

Ipadic vs Jumandic: There were major differences in numbers of each part-of-speech type annotations between the dictionaries. In most cases ipadic provided more specific annotations (nouns, verbs, particles, auxiliary verbs, exclamations) than jumandic. For example, in ipadic annotation there were nearly 2 billion of nouns, while in jumandic only about 1,5 billion (see table 3 and its graphical visualization in figure 3 for details). The differences are clearly visible when the category "other" is compared, which consists of such annotations as "symbols", or "unknown words". The number of "other" annotations with jumandic is over two times larger than with ipadic and covers nearly 40% of the whole corpus. The detailed analysis also revealed more generic differences in word coverage of the dictionaries. Especially when it comes to abbreviations and casual modifications, some words do not appear in jumandic. For example, an interjection いや *iya* ("oh") appears in both, but its casual modification いやー *iyaa* ("ooh") appears only in ipadic. In this situation jumandic splits the word in two parts: いや and a vowel prolongation mark ー, which is annotated by jumandic as "symbol".

YACIS vs jBlogs and JENAAD: It is difficult to manually evaluate annotations on a corpus as large as YACIS. However, the larger the corpus is the more statistically reliable are the observable tendencies of annotated phenomena. Therefore it is possible to evaluate the accurateness of annotations by comparing tendencies between different corpora. To verify part-of-speech tagging we compared tendencies in annotations between YACIS, jBlogs [7] and JENAAD [12]. The former, developed by Baroni and Ueyama [7], is a medium-sized corpus of Japanese blogs containing 62 million words. The corpus is based on four popular blog services (Ameba, Goo, Livedoor, Yahoo!). It contains nearly 30 thousand blog documents. The part of speech tagging was done by ChaSen. The latter is a medium-scale corpus of newspaper articles gathered from the Yomiuri daily newspaper (years 1989-2001). It contains about 4.7 million words (approximately 7% of jBlogs and 0.08% of YACIS). The comparison of those corpora provided interesting observations. jBlogs and JENAAD were annotated with ChaSen, while YACIS with MeCab. However, as mentioned in section 3, ChaSen and MeCab in their default settings use the same ipadic dictionary. Although there are some differences in the way each system disambiguates parts of speech, the same dictionary base makes it a good comparison of ipadic annotations on three different corpora (small JENAAD, larger jBlogs and large YACIS). The statistics of parts-of-speech distribution is more similar between the pair YACIS(ipadic)–JENAAD (ρ = 1.0 in Spearman's rank setting correlation test) and YACIS(ipadic)–jBlogs (ρ = 0.96), than between the pairs YACIS(jumandic)–jBlogs (ρ = 0.79), YACIS(jumandic)–JENAAD (ρ = 0.85) and between both version of YACIS (ρ = 0.88). See table 3 for details.

Japanese vs British and Italian: As an additional exercise we compared YACIS to Web corpora in different languages, namely ukWaC (British English) and itWaC (Italian) [13]. Although the information on part-of-speech distribution for those two corpora is incomplete, the available information shows interesting differences between languages[6]. In all compared corpora the largest is the

[5] http://nlp.cs.nyu.edu/irex/index-e.html

[6] We do not get into a discussion on differences between POS taggers for different languages, neither the discussion on whether the same POS names (like noun, verb, or adjective) represent similar concepts among different languages (see for example [14] or [15]). These two discussions, although important, are beyond the scope of this paper.

Table 3: Comparison of parts of speech distribution across corpora (percentage).

Part of speech	YACIS-ipadic		YACIS-jumandic		jBlogs	JENAAD	ukWaC	itWaC
	percentage	(number)	percentage	(number)	(approx.)	(approx.)		
Noun	34.69%	(1,942,930,102)	25.35%	(1,419,508,028)	34%	43%	1,528,839	941,990
Particle	23.31%	(1,305,329,099)	19.14%	(1,072,116,901)	18%	26%	[not provided]	[not provided]
Verb	11.57%	(647,981,102)	9.80%	(549,048,400)	9%	11%	182,610	679,758
Auxiliary verb	9.77%	(547,166,965)	2.07%	(115,763,099)	7%	5%	[not provided]	[not provided]
Adjective	2.07%	(116,069,592)	3.70%	(207,170,917)	2%	1%	538,664	706,330
Interjection	0.56%	(31,115,929)	0.40%	(22,096,949)	<1%	<1%	[not provided]	[not provided]
Other	18.03%	(1,010,004,306)	39.55%	(2,214,892,801)	29%	14%	[not provided]	[not provided]

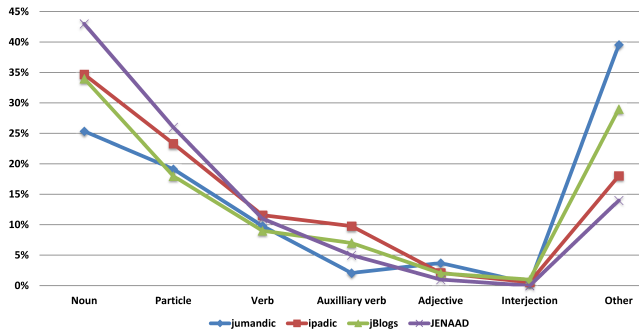

Figure 3: Graphical visualization of parts-of-speech comparison between YACIS (ipadic and jumandic annotations), Baroni&Ueyama's jBlogs and JENAAD.

number of "nouns". However, differently to all Japanese corpora, second frequent part of speech in British English and Italian was "adjective", while in Japanese it was "verb" (excluding particles). This difference is the most vivid in ukWaC. Further analysis of this phenomenon could contribute to the fields of language anthropology, and philosophy of language in general.

5 Conclusions and Future Work

In this paper we presented our research in annotating YACIS, a large corpus of Japanese blogs, with syntactic information. The information we annotated included tokenization, parts of speech, lemmatization, dependency structure, or named entities. In the annotation process we used standard tools for annotating these kinds of information on the Japanese language. The annotated corpus was compared to two other corpora in Japanese, and additionally to two corpora in different languages (British English and Italian). The comparison revealed interesting observations. The three corpora in Japanese, although different in size, showed similar POS distribution, whereas for other languages, although the corpora were comparable in size, the POS distribution differed greatly. We plan to address these differences in more detail in the future. In the near future we also plan to provide a demo viewable online allowing corpus querying for all types of annotations.

Acknowledgments

This research was supported by (JSPS) KAKENHI Grant-in-Aid for JSPS Fellows (Project Number: 22-00358).

References

[1] Erjavec, I. S., Erjavec, T., Kilgarriff, A. 2008. "A web corpus and word sketches for Japanese", *Information and Media Technologies*, 3(3), pp. 529-551.

[2] Aozora Bunko, http://www.aozora.gr.jp/ (Retrieved: 2012.01.25)

[3] Mainichi Shinbun CD, http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html (Retrieved: 2012.01.25)

[4] Abe, S., Eguchi, M., Sumida, A., Ohsaki, A., Inui, K. 2009. "Minna no keiken: Burogu kara chuushutsu shita ibento oyobi senchimento no DB-ka" (Everyone's experiences: Creating a Database of Events and Sentiments Extracted from Blogs) [in Japanese], In *Proceedings of NLP2009*, pp. 296-299.

[5] Hashimoto, C., Kurohashi, S., Kawahara, D., Shinzato, K. and Nagata, M. 2011. "Construction of a Blog Corpus with Syntactic, Anaphoric, and Sentiment Annotations" [in Japanese], *Journal of Natural Language Processing*, Vol. 18, No. 2, pp. 175-201.

[6] Quan, C. and Ren, F. 2010. "A blog emotion corpus for emotional expression analysis in Chinese", *Computer Speech & Language*, Vol. 24, Issue 4, pp. 726-749.

[7] Baroni, M. and Ueyama, M. 2006. "Building General- and Special-Purpose Corpora by Web Crawling", In *Proceedings of the 13th NIJL International Symposium on Language Corpora: Their Compilation and Application*.

[8] Yoshihira, K., Takeda, T., Sekine, S. 2004. "KWIC system for Web Documents" [in Japanese]. In *Proceedings of the 10th Annual Meetings of the Japanese Association for NLP*, pp. 137-139.

[9] Maciejewski, J., Ptaszynski, M., Dybala, P. 2010. "Developing a Large-Scale Corpus for Natural Language Processing and Emotion Processing Research in Japanese", In *Proceedings of the International Workshop on Modern Science and Technology (IWMST)*, pp. 192-195.

[10] Kudo, T. "MeCab: Yet Another Part of Speech and Morphological analyzer", http://mecab.sourceforge.net/ (Retrieved: 2012.01.25)

[11] Kudo, T. and Matsumoto, Y. 2002. "Japanese Dependency Analysis using Cascaded Chunking", In *Proceedings of the 6th Conference on Natural Language Learning 2002 (CoNLL 2002)*, pp. 63-69.

[12] Utiyama, M. and Isahara, H. 2003. "Reliable Measures for Aligning Japanese-English News Articles and Sentences", *ACL-2003*, pp. 72-79.

[13] Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E. 2008. "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora", Kluwer Academic Publishers, Netherlands.

[14] Hopper, P. and Thompson, S. 1985. "The Iconicity of the Universal Categories 'Noun' and 'Verbs'". In *Typological Studies in Language: Iconicity and Syntax*. John Haiman (ed.), vol. 6, pp. 151-183, Amsterdam, John Benjamins Publishing Company.

[15] Broschart, J. 1997. "Why Tongan does it differently: Categorial Distinctions in a Language without Nouns and Verbs", *Linguistic Typology*, Vol. 1, No. 2, 123-165.