A Japanese Word Dependency Corpus

Shinsuke Mori,[1] Hideki Ogura,[2] Tetsuro Sasada[3]
[1,3]Academic Center for Computing and Media Studies, Kyoto University
[2]College of Letters, Ritsumeikan University

1 Introduction

Empirical methodology in natural language processing (NLP) has experienced a great success recently [1]. In this methodology, language resource availability is the most important. In fact Penn Treebank [2], annotated with part-of-speech (POS) information and phrase structure, has driven various researches on POS tagging models and parsing models. Therefore similar corpora have been developed in various languages.

For Japanese, the language we focus on this paper, a high quality balanced corpus called Balanced Corpus of Contemporary Written Japanese (BCCWJ) [3] has been issued recently. The sources of the corpus varies from newspaper articles to blogs. The sentences in the core part of this corpus (BCCWJ Core data) are segmented into words, called a "short-unit word" and each word is annotated with a part-of-speech (POS) and pronunciation. Using the BCCWJ as a training corpus, high accuracy has been achieved in word segmentation [4], POS tagging [5], and pronunciation estimation [6]. This corpus, however, does not have other higher order phenomena.

Among these we focused on the dependency structure and we annotated more than 30 thousands sentences including the BCCWJ Core data. For Japanese there is a dependency corpus [7]. Its unit is, however, phrase called *bunsetsu*. This Japanese specific unit is not compatible with the word unit in other languages. Thus there is a strong requirement of a word-based dependency corpus.

In this background, first we designed the standard of word dependency annotation for Japanese. The word unit is compatible with BCCWJ. Next we annotated more than 30 thousand sentences with dependency structure. In this paper, we first present the specification of our corpus. Then we give a detailed explanation about our standard of word dependency. We also report some experiments on dependency parsing using our corpus.

2 Corpus Specification

In this section, we present the details of our word dependency corpus, except for the dependency standard, which we discuss in the next section.

2.1 Unit Definition

For the dependency annotation unit, we have chosen the word as in many languages. As we noted in the previous section, a language specific unit called *bunsetsu* is famous for Japanese dependency description [7]. This unit is, however, too long for various applications. In fact in some languages, a sentence is separated into phrases by white spaces when it is written[1]. But phrases are divided into some smaller units in many researches [11]. From the above observation, we decided to take word as the unit of our dependency corpus.

For the definition of word, we follow that of BCCWJ, which is a mature standard created by linguists of Japanese language. The only difference is that we separate the endings of inflectional words (adjectives, verbs, and auxiliary verbs) from their stems for two reasons.

1. By taking stems and endings into the vocabulary separately, we can build a higher coverage language model (LM) with a smaller vocabulary. This allows us to increase the performance of LM-based applications such as an automatic speech recognizer (ASR) [12] and input method (IM) [13].
2. By separating endings from stems, we can identify different inflection forms of the same verb just by a string match[2]. That is, we do not need to prepare the list of inflection patterns and the correct inflection pattern at the step of morphological analysis as well.

2.2 Source and Size

Some experimental results [14] demonstrate that the parsing accuracy is high enough for real applications if a high quality dependency corpus is available in the application domain. Now the focus of parsing research has been shifting to domain adaptability of methods. Therefore, we decided to take sentences from various domains to allow corpus users to conduct domain adaptation experiments. Table 1 shows specifications of our corpus. Each word, except for the root word, is annotated with its head (dependency destination). Thus the number of dependencies in a corpus is equal to the number of words minus the number of sentences.

Below we explain the features of each domain and the reason why we have chosen them.

2.2.1 BCCWJ Core data

BCCWJ [3] has a core part whose sentences are manually segmented into words and the words are annotated with their POS and pronunciation. The an-

[1] In many researches on these languages, these phrases are called word because of they are visually similar to English word but they are phrase in granularity of meaning.

[2] Some words such as "行く" (go) and "行う" (execute) share the stem ("行" in these examples). This ambiguity may be resolved by a method for word sense disambiguation.

Table 1: Corpus specifications.

ID		source	#Sentences	#Words	#Characters
BCCWJ	OC	Yahoo! questions and answers	615	12,487	17,294
	OW	White papers	658	26,546	38,847
	OY	Yahoo! blog	857	13,386	19,833
	PB	Books	1,058	23,473	32,356
	PM	Magazines	1,505	25,274	39,842
	PN	Newspaper articles	1,713	38,063	55,454
		subtotal	6,406	139,229	203,626
EHJ		Dictionary example sentences	13,000	162,273	220,148
NKN		Economy newspaper articles	10,025	292,253	442,264
NPT		NTCIR patent disclosure	500	20,653	32,139
		total	29,931	614,408	898,177

For the latest specifications see `http://plata.ar.media.kyoto-u.ac.jp/tool/EDA/model.html` .

notation quality is very high and the accuracies of POS tagging and that of pronunciation estimation are both more than 98%.

We annotated 1/10 of this part with word dependency. These data allow NLP researchers to work on joint models for POS tagging and dependency parsing [15, 9] and structured language models [16, 17] for automatic speech recognition [12] or input methods [13]. A research on the influence of syntactic structure to the pronunciation is also interesting since the pronunciation estimation of some important words can only be solved by referring to long dependencies.

Some researchers are annotating BCCWJ Core data about other linguistic phenomena including predicate-argument structure, coreference, etc. With our dependency annotation, various researches are expected to be possible.

2.2.2 Dictionary example sentences: EHJ

We annotated about 80% sentences of the example sentences in a dictionary for daily conversation [18]. There are two important features. The first one is that this set covers the basic vocabulary in Japanese consisting of about 2,500 words in various basic meanings. Our dependency annotation on this set is useful to build a parser for spoken Japanese. The second feature is that each Japanese sentence has its English translation[3], which is useful for machine translation (MT) experiments.

The sentences have word boundary information of course. And words are annotated with their pronunciation but not with POS tag. We conducted an experiment of automatic word segmentation and POS tagging. The result showed that a publicly available state-of-the-art POS tagger *KyTea* [5] trained on BCCWJ achieved about 98% accuracy on a small subset of these sentences.

2.2.3 Economy newspaper articles: NKN

Penn Treebank [2] consists of sentences in Wall Street Journal, which is a newspaper for economy. So we focused on a newspaper specialized in economy. In Japanese *Nikkei* newspaper is the only clear counterpart of Wall Street Journal. We annotated the sentences taken from this newspaper with word boundary information and dependency structure. This allows researchers to compare Japanese and English.

BCCWJ has a subset taken from articles of general newspapers (PN in Table 1). However, Table 1 indicates that the average sentence length of this *Nikkei* set is 29.2 words which is much larger than that of BCCWJ PN, the second longest set (22.2 words).

Similar to EHJ, words are annotated with their pronunciation but not with POS tag. An experiment of word segmentation and POS tagging in the same setting as the EHJ case showed that the accuracy is about 96%.

2.2.4 Invention disclosures: NPT

NTCIR deploys a shared task for patent machine translation [10] and makes English-Japanese sentence pairs taken from invention disclosures publicly available. We annotated a small part of this set with word boundary information and dependency structure.

With this set we can adapt a dependency parser to the patent domain and measure the parsing accuracy. Then MT researchers can use that parser to automatically annotate invention disclosure sentences with dependency structure and work on tree-based machine translation.

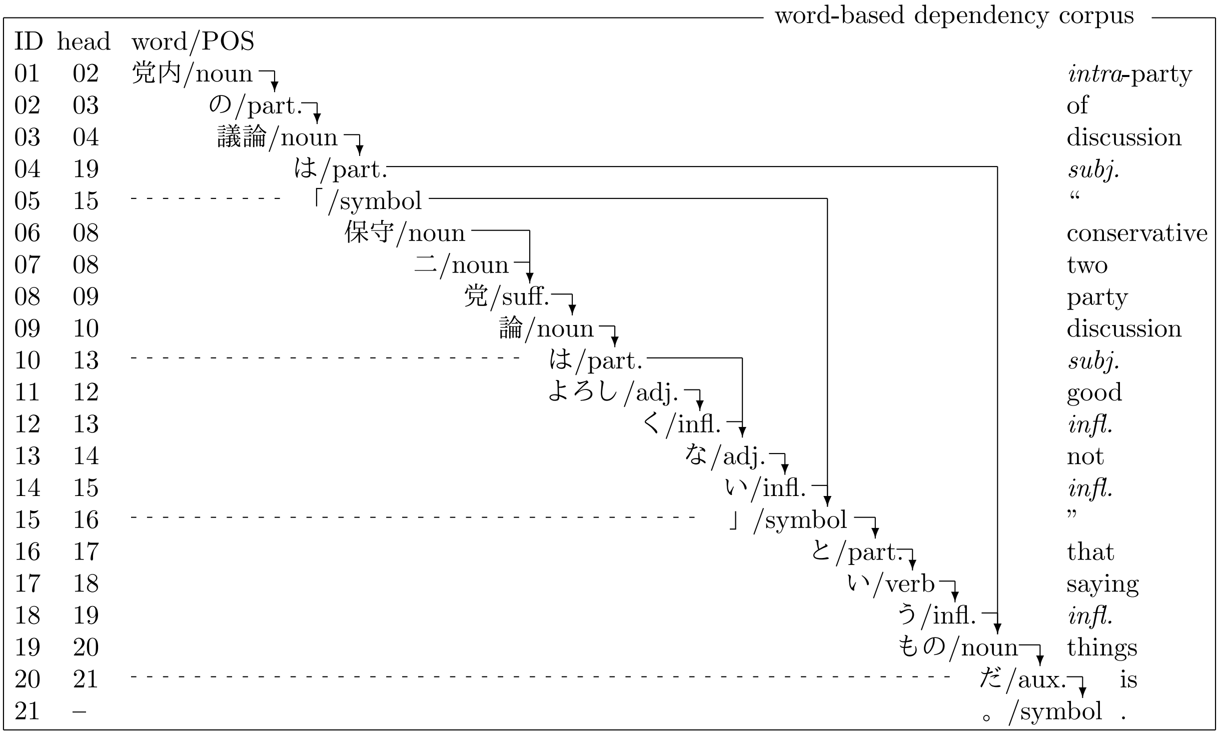
3 Annotation Standard

The dependency annotation standard of our corpus is basically similar to that of other treebanks. That is to say, a source word w_s depends on another word w_h, called a head, that the word modifies and the concatenation of the source word and the head $w_s w_h$ should be a natural word sequence which may appear in a huge corpus. Figure 1 shows an example. In this section we present regulations for frequent phenomena taken from our annotation guideline.

3.1 Simple sentence

Basically Japanese is an SOV language. That is to say, the word order in a simple sentence is subject, object, and verb. Almost all noun phrases have a

[3] The French and German translation is also available in printed version but not in machine readable form.



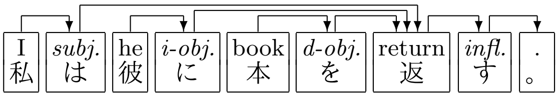
word-based dependency corpus

ID	head	word/POS	
01	02	党内/noun	*intra*-party
02	03	の/part.	of
03	04	議論/noun	discussion
04	19	は/part.	*subj.*
05	15	「/symbol	“
06	08	保守/noun	conservative
07	08	二/noun	two
08	09	党/suff.	party
09	10	論/noun	discussion
10	13	は/part.	*subj.*
11	12	よろし/adj.	good
12	13	く/infl.	*infl.*
13	14	な/adj.	not
14	15	い/infl.	*infl.*
15	16	」/symbol	”
16	17	と/part.	that
17	18	い/verb	saying
18	19	う/infl.	*infl.*
19	20	もの/noun	things
20	21	だ/aux.	is
21	–	。/symbol	.

Figure 1: An example of dependencies for a sentence.

case marker called postposition to clarify its role to the verb. The only limitation is to put the main verb phrase at the end. That is to say, subject (*subj.*), direct object (*d-obj.*), indirect object (*i-obj.*), and other verb modifier such as adverbial phrases are ordered freely.

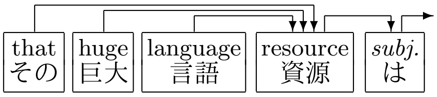
In our corpus, the head of a noun phrase w_n depends on its postposition w_p, and w_p depends on the verb w_v as shown in the example below.



3.2 Compound word

We annotate a compound word with the structure representing its meaning. Modifiers of a compound word depend on its head (in many cases with very few exceptions which modifies a part of a compound word) and there is only one dependency arc going out from the head.

Let us take a noun phrase example, “huge language resource.”



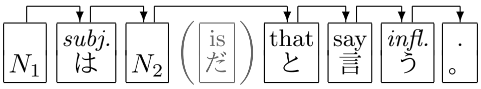
In this example “huge” depends on “resource” because what is “huge” is not “language” but “resource.” Another modifier, “that,” depends on the head of the noun phrase, “resource,” and it depends on the following postposition.

3.3 Copula

Some sentences have a copular verb. Most copula sentences fall into the following type.

N_1 は/*subj.* N_2 だ/is

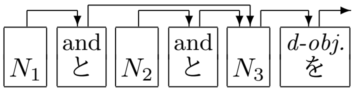
We decided that the case marker “は/*subj.*” depends on N_2, not on the auxiliary verb “だ/is.” The reason is that an auxiliary verb can be omitted especially in a *that*-clause or sentence coordination. The head of the case marker is always N_2 independent from the existence of an auxiliary verb.



This is somewhat debatable because this breaks the structural compatibility with many European languages and makes the tree-based machine translation complicated.

3.4 Coordination

A coordination structure is also a frequent phenomenon. In a coordination structure two or more phrases are concatenated by using a coordination marker. In Japanese the most frequent marker is “と/and.” This marker is similar to “and” in English but we put one at each point between elements as follows.



In this case, our annotation standard states that N_1 and N_2 depend on each marker following them. The markers depends on the last element N_3, not on the next element.

4 Parsing Experiments

The most typical usage of our corpus is to build a parser. In this section, we present parsing experiment results on our corpus.

Table 2: Parsing accuracy.

		#Sentences		
ID		Training	Test	Accuracy
BCCWJ	OC	365	250	95.11%
	OW	408	250	91.27%
	OY	607	250	89.63%
	PB	808	250	94.14%
	PM	1,255	250	95.80%
	PN	1,463	250	92.66%
EHJ		11,700	1,300	97.07%
NKN		9,023	1,002	93.22%
NPT		450	50	90.92%

The parser we used is MST-based dependency parser *EDA*[4] [19]. We divided all the subset into a training and a test part (see Table 2). Then we build a single model of *EDA* from all the training sets and measured the word-based accuracy on each test set.

From the results shown in Table 2, it can be said that the easiest is the set of dictionary example sentences (EHJ). Magazines (BCCWJ PM) and Yahoo! questions and answers (BCCWJ OC) are the second easiest. The reason may be their limitation on the vocabulary and sentence pattern variations. The most difficult is the blog domain (BCCWJ OY). This set is composed of user generated contents (UGC) and its topic varies widely. The invention disclosure set (NPT) is also difficult. The sentences tend to be long and the writing style is different. There is, however, a clear application for this set, which is tree-based machine translation. We need more training data to increase the accuracy in these domains.

5 Conclusion

In this paper, we reported the details of our word-based dependency corpus in Japanese. The unit is compatible with the Balanced Corpus of Contemporary Written Japanese (BCCWJ), which is of high quality and widely used for various NLP tasks. The size of our corpus is about 30 thousand sentences, which is enough to train statistical parsers for the general domain. We then discussed the dependency annotation standard, and finally reported some preliminary results of an MST-based dependency parser on our corpus.

Acknowledgments

This work was supported by JSPS Grants-in-Aid for Scientific Research Grant Numbers 23500177, NTT agreement dated 05/23/2013, and Basic Research on Corpus Annotation project of The National Institute for Japanese Language and Linguistics. We are also grateful to the annotators for their contribution to the design of the guidelines and the annotation effort.

[4] `http://plata.ar.media.kyoto-u.ac.jp/tool/EDA/home_en.html` (accessed on 2014/Feb./01).

References

[1] Armstrong, S.(ed.): *Using Large Corpora*, The MIT Press (1994).

[2] Marcus, M. P. and Santorini, B.: Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, Vol. 19, No. 2, pp. 313–330 (1993).

[3] Maekawa, K., Yamazaki, M., Maruyama, T., Yamaguchi, M., Ogura, H., Kashino, W., Ogiso, T., Koiso, H. and Den, Y.: Design, Compilation, and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese, *Proc. of the LREC10* (2010).

[4] Neubig, G. and Mori, S.: Word-based Partial Annotation for Efficient Corpus Construction, *Proc. of the LREC10* (2010).

[5] Neubig, G., Nakata, Y. and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, *Proc. of the ACL11*, pp. 529–533 (2011).

[6] Mori, S. and Neubig, G.: A Pointwise Approach to Pronunciation Estimation for a TTS Front-end, *Proc. of the InterSpeech2011*, Florence, Italy, pp. 2181–2184 (2011).

[7] Kurohashi, S. and Nagao, M.: Building a Japanese Parsed Corpus while Improving the Parsing System, *Proc. of the LREC98*, pp. 719–724 (1998).

[8] Buchholz, S. and Marsi, E.: CoNLL-X shared task on Multilingual Dependency Parsing, *Proc. of the CoNLL2006*, pp. 149–164 (2006).

[9] Hatori, J., Takuya, M., Yusuke, M. and Jun'ichi, T.: Incremental Joint POS Tagging and Dependency Parsing in Chinese, *Proc. of the IJCNLP11* (2011).

[10] Goto, I., Lu, B., Chow, K. P., Sumita, E. and Tsou, B. K.: Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop, *Proceedings of NTCIR-9 Workshop Meeting*, pp. 559–578 (2011).

[11] Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S. and Pylkkönen, J.: Unlimited vocabulary speech recognition with morph language models applied to Finnish, *Computer Speech and Language*, Vol. 20, pp. 515–541 (2006).

[12] Bahl, L. R., Jelineck, F. and Mercer, R. L.: A Maximum Likelihood Approach to Continuous Speech Recognition, *IEEE PAMI*, Vol. 6, No. 2, pp. 179–190 (1983).

[13] Mori, S., Takuma, D. and Kurata, G.: Phoneme-to-Text Transcription System with an Infinite Vocabulary, *Proc. of the COLING06* (2006).

[14] McDonald, R. and Nivre, J.: Analyzing and Integrating Dependency Parsers, *Computational Linguistics*, Vol. 37, No. 4, pp. 197–230 (2011).

[15] Mori, S., Nishimura, M., Itoh, N., Ogino, S. and Watanabe, H.: A Stochastic Parser Based on a Structural Word Prediction Model, *Proc. of the COLING00*, pp. 558–564 (2000).

[16] Chelba, C. and Jelinek, F.: Structured Language Modeling, *Computer Speech and Language*, Vol. 14, pp. 283–332 (2000).

[17] Mori, S., Nishimura, M. and Itoh, N.: Improvement of a Structured Language Model: Arbori-context Tree, *Proc. of the EuroSpeech2001* (2001).

[18] Keene, D., Hatori, H., Yamada, H. and Irabu, S.: *Japanese-English Sentence Equivalents*, Asahi Press, Electronic book edition (1992).

[19] Flannery, D., Miyao, Y., Neubig, G. and Mori, S.: Training Dependency Parsers from Partially Annotated Corpora, *Proc. of the IJCNLP11* (2011).

[20] McDonald, R., Pereira, F., Ribarov, K. and Hajič, J.: Non-projective Dependency Parsing Using Spanning Tree Algorithms, *Proc. of the EMNLP*, pp. 523–530 (2005).