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]Faculty of Literature, Ritsumeikan University
[1,3]Yoshidahonmachi, Sakyo-ku, Kyoto, Japan
[2]Toujiinkitamachi, Kita-ku, Kyoto, Japan
[1]forest@i.kyoto-u.ac.jp, [2]h-ogura@fc.ritsumei.ac.jp, [3]sasada@ar.media.kyoto-u.ac.jp

Abstract

In this paper, we present a corpus annotated with dependency relationships in Japanese. It contains about 30 thousand sentences in various domains. Six domains in Balanced Corpus of Contemporary Written Japanese have part-of-speech and pronunciation annotation as well. Dictionary example sentences have pronunciation annotation and cover basic vocabulary in Japanese with English sentence equivalent. Economic newspaper articles also have pronunciation annotation and the topics are similar to those of Penn Treebank. Invention disclosures do not have other annotation, but it has a clear application, machine translation. The unit of our corpus is word like other languages contrary to existing Japanese corpora whose unit is phrase called *bunsetsu*. Each sentence is manually segmented into words. We first present the specification of our corpus. Then we give a detailed explanation about our standard of word dependency. We also report some preliminary results of an MST-based dependency parser on our corpus.

Keywords: Dependency corpus, Word unit, Japanese

1. Introduction

Empirical methodology in natural language processing (NLP) has experienced a great success recently (Armstrong, 1994). In this methodology, language resource availability is the most important. In fact Penn Treebank (Marcus and Santorini, 1993), 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) (Maekawa et al., 2010) 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 (Neubig and Mori, 2010), POS tagging (Neubig et al., 2011), and pronunciation estimation (Mori and Neubig, 2011). This corpus, however, does not have syntactic information or other higher order phenomena. Thus, there are some attempts at adding syntactic structure, predicate-argument structure, coreferences, etc. to the BCCWJ Core data.

Among these we focused on the dependency structure (the next step of NLP after POS tagging) and we annotated more than 30 thousands sentences including the BCCWJ Core data. For Japanese there is a dependency corpus (Kurohashi and Nagao, 1998). Its unit is, however, phrase called *bunsetsu*, which consists of one or more content words and zero or more function words. 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 CoNLL shared task on dependency parsing (Buchholz and Marsi, 2006) for example, Japanese corpus is the automatic conversion result of *bunsetsu*-based dependency, where they decided that dependencies among words in a *bunsetsu* are always left-to-right and the last word of each *bunsetsu* is always its head ignoring the real structure of compound words, head-and-modifier relationships.

In this background, first we originally 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. The annotated data are composed of the following sources:

BCCWJ Core data, which allows us to work on a joint model for POS tagging and dependency parsing (Hatori et al., 2011) or for more complicated linguistic phenomena if these annotations are issued from other sites,

dictionary example sentences, which cover the basic vocabulary in Japanese and have the English translations,

economy newspaper articles, which are similar to Wall Street Journal of Penn Treebank and allows us to compare the result with it,

invention disclosures, which were taken from the NTCIR patent translation task (Goto et al., 2011) and thus enables tree-based machine translation.

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.

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

Table 1: Corpus specifications

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 (Kurohashi and Nagao, 1998). 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 (Hirsimäki et al., 2006). 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) (Bahl et al., 1983) and input method (IM) (Mori et al., 2006).

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.

Note that the BCCWJ original "short-unit word" dependency can be trivially obtained just by concatenating the stems and the inflectional endings and then erasing the dependencies between them.

2.2. Source and Size

Some experimental results (McDonald and Nivre, 2011) 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 (Maekawa et al., 2010) has a core part whose sentences are manually segmented into words and the words are annotated with their POS and pronunciation. The annotation 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 (Mori et al., 2000; Hatori et al., 2011) and structured language models (Chelba and Jelinek, 2000; Mori et al., 2001) for automatic speech recognition (Bahl et al., 1983) or input methods (Mori et al., 2006). 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 (Keene et al., 1992). 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

[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.

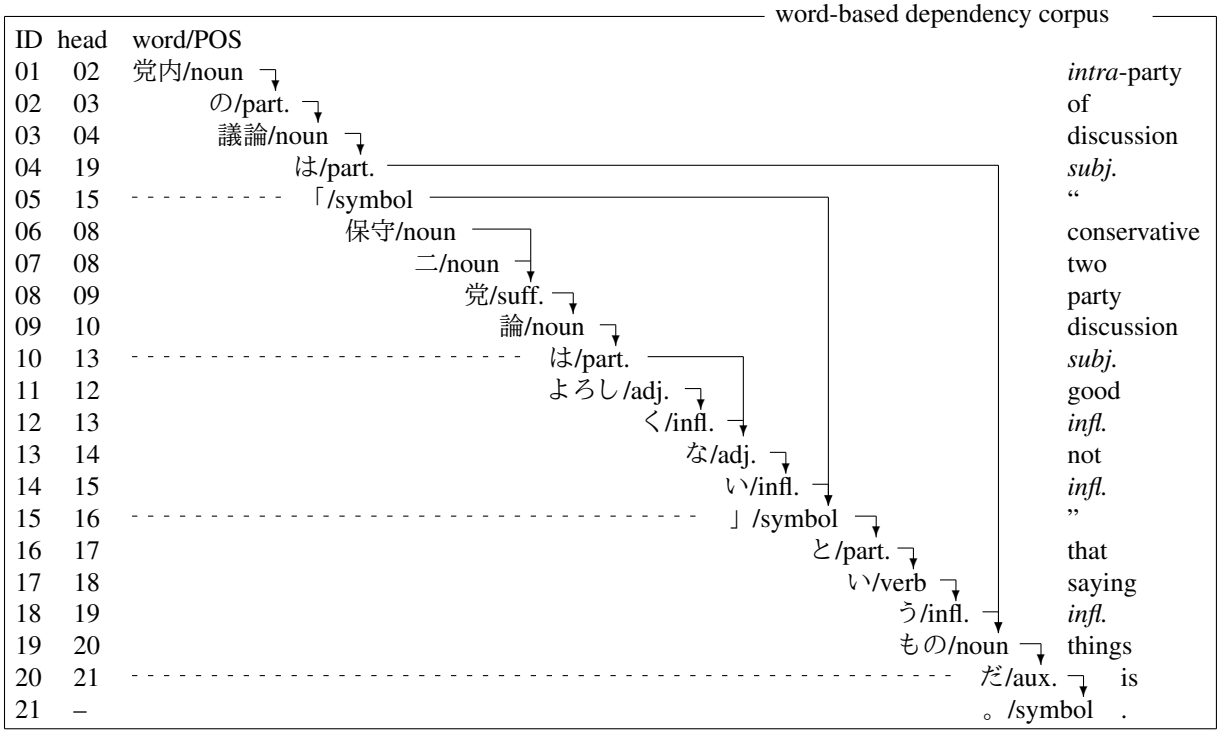


Figure 1: An example of dependencies for a sentence.

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* (Neubig et al., 2011) trained on BCCWJ achieved about 98% accuracy on a small subset of these sentences.

2.2.3. Economy newspaper articles: NKN

Penn Treebank (Marcus and Santorini, 1993) 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%.

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

2.2.4. Invention disclosures: NPT

NTCIR deploys a shared task for patent machine translation (Goto et al., 2011) 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. Dependency Annotation Standard

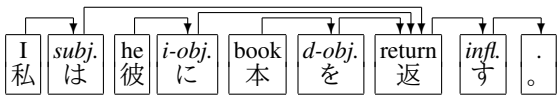
The dependency annotation standard of our corpus is basically similar to that of other treebanks. That is to say, a source word w_s depends on another word w_h, called a head, that the word modifies and the concatenation of the source word and the head $w_s w_h$ should be a natural word sequence which 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 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.

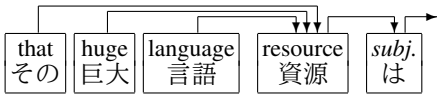
In our corpus, the head of a noun phrase w_n depends on its postposition w_p, and w_p depends on the verb w_v as shown

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."



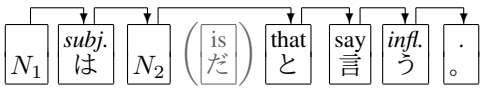
In this example "huge" depends on "resource" because what is "huge" is not "language" but "resource." Another modifier, "that," depends on the head of the noun phrase, "resource," and it depends on the following postposition.

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

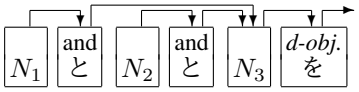
We decided that the case marker "は/*subj.*" depends on N_2, not on the auxiliary verb "だ/is." The reason is that an auxiliary verb can be omitted especially in a *that*-clause or sentence coordination. The head of the case marker is always N_2 independent from the existence of an auxiliary verb.



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.

ID		#Sentences 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%

Table 2: Parsing Accuracy.

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.

4.1. Experimental Settings

The parser we used is MST-based dependency parser *EDA*[4] (Flannery et al., 2011). 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.

4.2. A Pointwise MST parser

The parser *EDA* follows the standard setting of recent work on dependency parsing (Buchholz and Marsi, 2006). It assumes a sequence of words, $\boldsymbol{w} = \langle w_1, w_2, \ldots, w_n \rangle$, as an input and outputs a dependency tree $\boldsymbol{d} = \langle d_1, d_2, \ldots, d_n \rangle$, where $d_i \equiv j$ when the head of w_i is w_j. And $d_i = 0$ for some word w_i in a sentence, which indicates that w_i is the head of the sentence.
Like many other parsers, *EDA* is based on the maximum spanning tree (MST) algorithm (McDonald et al., 2005; McDonald and Nivre, 2011), in which a node corresponds to a word and a score, $\sigma(d_i)$, is assigned to each edge (i.e. dependency) d_i, and parsing finds a dependency tree, $\hat{\boldsymbol{d}}$, that maximizes the sum of the scores of all the edges as follows:

$$\hat{\boldsymbol{d}} = \mathop{\mathbf{argmax}}_{\boldsymbol{d}} \sum_{d \in \boldsymbol{d}} \sigma(d). \quad (1)$$

To estimate $\sigma(d)$ from an annotated corpus, we calculate the probability of a dependency labeling $p(d_i = j)$ for a word w_i from its context, which is a tuple $x = \langle \boldsymbol{w}, \boldsymbol{t}, i \rangle$, where $\boldsymbol{t} = \langle t_1, t_2, \ldots, t_n \rangle$ is a sequence of POS tags assigned to $\boldsymbol{w}$ by a tagger *KyTea*[5] (Neubig et al., 2011). Thus $\sigma(d)$ is estimated as the conditional probability $p(j|x)$ given by the following equation:

$$p(j|x, \boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta} \cdot \boldsymbol{\phi}(x, j)\right)}{\sum_{j' \in \mathcal{J}} \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{\phi}(x, j')\right)}. \quad (2)$$

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

[5] http://www.phontron.com/kytea/ (accessed on 2014/Feb./01).

The feature vector $\boldsymbol{\phi} = \langle \phi_1, \phi_2, \ldots, \phi_m \rangle$ is a vector of non-negative values calculated from features on pairs (x, j), with corresponding weights given by the parameter vector $\boldsymbol{\theta} = \langle \theta_1, \theta_2, \ldots, \theta_m \rangle$. We estimate $\boldsymbol{\theta}$ from the summation of the training parts of all the domains of our dependency corpus. It should be noted that the probability $p(d_i)$ depends only on i, j, and the inputs $\boldsymbol{w}$, $\boldsymbol{t}$, which ensures that it is estimated independently for each w_i.

4.3. Parsing Results

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.

6. 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.

7. References

Susan Armstrong, editor. 1994. *Using Large Corpora*. The MIT Press.

Lalit R. Bahl, Frederick Jelineck, and Robert L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2):179–190.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164.

Ciprian Chelba and Frederick Jelinek. 2000. Structured language modeling. *Computer Speech and Language*, 14:283–332.

Daniel Flannery, Yusuke Miyao, Graham Neubig, and Shinsuke Mori. 2011. Training dependency parsers from partially annotated corpora. In *Proceedings of the Fifth International Joint Conference on Natural Language Processing*.

Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. 2011. Overview of the patent machine translation task at the NTCIR-9 workshop. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 559–578.

Jun Hatori, Matsuzaki Takuya, Miyao Yusuke, and Tsujii Jun'ichi. 2011. Incremental joint POS tagging and dependency parsing in Chinese. In *Proceedings of the Fifth International Joint Conference on Natural Language Processing*.

Teemu Hirsimäki, Mathias Creutz, Vesa Siivola, Mikko Kurimo, Sami Virpioja, and Janne Pylkkönen. 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech and Language*, 20:515–541.

Donald Keene, Hiroyoshi Hatori, Haruko Yamada, and Shouko Irabu. 1992. *Japanese-English Sentence Equivalents*. Asahi Press, Electronic book edition.

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.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. 2010. Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Mitchell P. Marcus and Beatrice Santorini. 1993. Building a large annotated corpus of English: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Ryan McDonald and Joakim Nivre. 2011. Analyzing and integrating dependency parsers. *Computational Linguistics*, 37(4):197–230.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 523–530.

Shinsuke Mori and Graham Neubig. 2011. A pointwise approach to pronunciation estimation for a TTS front-end. In *Proceedings of the InterSpeech2011*, pages 2181–2184, Florence, Italy.

Shinsuke Mori, Masafumi Nishimura, Nobuyasu Itoh, Shiho Ogino, and Hideo Watanabe. 2000. A stochastic parser based on a structural word prediction model. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 558–564.

Shinsuke Mori, Masafumi Nishimura, and Nobuyasu Itoh. 2001. Improvement of a structured language model: Arbori-context tree. In *Proceedings of the Seventh European Conference on Speech Communication and Tech-*

nology.

Shinsuke Mori, Daisuke Takuma, and Gakuto Kurata. 2006. Phoneme-to-text transcription system with an infinite vocabulary. In *Proceedings of the 21th International Conference on Computational Linguistics*.

Graham Neubig and Shinsuke Mori. 2010. Word-based partial annotation for efficient corpus construction. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.