Word Order Acquisition from Corpora

Kiyotaka Uchimoto[†], **Masaki Murata**[†], **Qing Ma**[†],
Satoshi Sekine[‡], and **Hitoshi Isahara**[†]

[†]Communications Research Laboratory
Ministry of Posts and Telecommunications
588-2, Iwaoka, Iwaoka-cho, Nishi-ku
Kobe, Hyogo, 651-2492, Japan
[uchimoto,murata,qma,isahara]@crl.go.jp

[‡]New York University
715 Broadway, 7th floor
New York, NY 10003, USA
sekine@cs.nyu.edu

Abstract

In this paper we describe a method of acquiring word order from corpora. Word order is defined as the order of modifiers, or the order of phrasal units called 'bunsetsu' which depend on the same modifiee. The method uses a model which automatically discovers what the tendency of the word order in Japanese is by using various kinds of information in and around the target bunsetsus. This model shows us to what extent each piece of information contributes to deciding the word order and which word order tends to be selected when several kinds of information conflict. The contribution rate of each piece of information in deciding word order is efficiently learned by a model within a maximum entropy framework. The performance of this trained model can be evaluated by checking how many instances of word order selected by the model agree with those in the original text. In this paper, we show that even a raw corpus that has not been tagged can be used to train the model, if it is first analyzed by a parser. This is possible because the word order of the text in the corpus is correct.

1 Introduction

Although it is said that word order is free in Japanese, linguistic research shows that there are certain word order tendencies — adverbs of time, for example, tend to precede subjects, and bunsetsus in a sentence that are modified by a long modifier tend to precede other bunsetsus in the sentence. Knowledge of these word order tendencies would be useful in analyzing and generating sentences.

In this paper we define *word order* as the order of modifiers, or the order of bunsetsus which depend on the same modifiee. There are several elements which contribute to deciding the word order, and they are summarized by Saeki (Saeki, 1998) as basic conditions that govern word order. When interpreting these conditions according to our definition, we can summarize them as follows.

Componential conditions

- A bunsetsu having a deep dependency tends to precede a bunsetsu having a shallow dependency.

  When there is a long distance between a modifier and its modifiee, the modifier is defined as a bunsetsu having a deep dependency. For example, the usual word order of modifiers in Japanese is the following: a bunsetsu which contains an interjection, a bunsetsu which contains an adverb of time, a bunsetsu which contains a subject, and a bunsetsu which contains an object. Here, the bunsetsu containing an adverb of time is defined as a bunsetsu having deeper dependency than the one containing a subject. We call the concept representing the distance between a modifier and its modifiee the *depth of dependency.*
- A bunsetsu having wide dependency tends to precede a bunsetsu having narrow dependency.

  A bunsetsu having wide dependency is defined as a bunsetsu which does not rigidly restrict its modifiee. For example, the bunsetsu "*Tokyo_e* (to Tokyo)" often depends on a bunsetsu which contains a verb of motion such as "*iku* (go)" while the bunsetsu "*watashi_ga* (I)" can depend on a bunsetsu which contains any kind of verb. Here, the bunsetsu "*watashi_ga* (I)" is defined as a bunsetsu having wider dependency than the bunsetsu "*Tokyo_e* (to Tokyo)." We call the concept of how rigidly a modifier restricts its modifiee the *width of dependency.*

Syntactic conditions

- A bunsetsu modified by a long modifier tends to precede a bunsetsu modified by a short modifier.

  A long modifier is a long clause, or a clause that contains many bunsetsus.
- A bunsetsu containing a reference pronoun tends to precede other bunsetsus in the sentence.
- A bunsetsu containing a repetition word tends to precede other bunsetsus in the sentence.

  A repetition word is a word referring to a word in a preceding sentence. For example, Taro and Hanako in the following text are repetition words. "Taro and Hanako love each other. Taro is a civil servant and Hanako is a doctor."
- A bunsetsu containing the case marker "wa" tends to precede other bunsetsus in the sentence.

A number of studies have tried to discover the relationship between these conditions and word order in

Japanese. Tokunaga and Tanaka proposed a model for estimating Japanese word order based on a dictionary. They focused on the width of dependency (Tokunaga and Tanaka, 1991). Under their model, however, word order is restricted to the order of case elements of verbs, and it is pointed out that the model can deal with only the obligatory case and it cannot deal with contextual information (Saeki, 1998). An N-gram model for detecting word order has also been proposed by Maruyama (Maruyama, 1994), but under this model word order is defined as the order of morphemes in a sentence. The problem setting of Maruyama's study thus differed from ours, and the conditions listed above were not taken into account in that study. As for estimating word order in English, a statistical model has been proposed by Shaw and Hatzivassiloglou (Shaw and Hatzivassiloglou, 1999). Under their model, however, word order is restricted to the order of premodifiers or modifiers depending on nouns, and the model does not simultaneously take into account many elements that contribute to determining word order. It would be difficult to apply the model to estimating word order in Japanese when considering the many conditions as listed above.

In this paper, we propose a method for acquiring from corpora the relationship between the conditions itemized above and word order in Japanese. The method uses a model which automatically discovers what the tendency of the word order in Japanese is by using various kinds of information in and around the target bunsetsus. This model shows us to what extent each piece of information contributes to deciding the word order and which word order tends to be selected when several kinds of information conflict. The contribution rate of each piece of information in deciding word order is efficiently learned by a model within a maximum entropy (M.E.) framework. The performance of the trained model can be evaluated according to how many instances of word order selected by the model agree with those in the original text. Because the word order of the text in the corpus is correct, the model can be trained using a raw corpus instead of a tagged corpus, if it is first analyzed by a parser. In this paper, we show experimental results demonstrating that this is indeed possible even when the parser is only 90% accurate.

This work is a part of the corpus based text generation. A whole sentence can be generated in the natural order by using the trained model, given dependencies between bunsetsus. It could be helpful for several applications such as refinement support and text generation in machine translation.

2 Word Order Acquisition and Estimation

2.1 Word Order Model

This section describes a model which estimates the likelihood of the appropriate word order. We call this model a *word order model*, and we implemented it within an M.E. framework.

Given tokenization of a test corpus, the problem of word order estimation in Japanese can be reduced to the problem of assigning one of two tags to each relationship between two modifiers. A relationship could be tagged with "1" to indicate that the order of the two modifiers is appropriate, or with "0" to indicate that it is not. Ordering all modifiers so as to assign the tag "1" to all relationships indicates that all modifiers are in the appropriate word order. The two tags form the space of "futures" in the M.E. formulation of our estimation problem of word order between two modifiers. The M.E. model, as well as other similar models allows the computation of $P(f|h)$ for any f in the space of possible futures, F, and for every h in the space of possible histories, H. A "history" in maximum entropy is all of the conditioning data that enable us to make a decision in the space of futures. In the estimation problem of word order, we could reformulate this in terms of finding the probability of f associated with the relationship at index t in the test corpus as:

$$P(f|h_t) = P(f|\text{ Information derivable from the test corpus related to relationship } t)$$

The computation of $P(f|h)$ in any M.E. models is dependent on a set of "features" which should be helpful in making a prediction about the future. Like most current M.E. models in computational linguistics, our model is restricted to features which are binary functions of the history and future. For instance, one of our features is

$$g(h,f) = \begin{cases} 1 : \text{if has}(h,x) = \text{true}, \\ \quad x = \text{“Mdfr1 − Head−} \\ \quad\quad \text{POS(Major) : verb''} \\ \quad \& \ f = 1 \\ 0 : \text{otherwise.} \end{cases} \quad (1)$$

Here "has(h,x)" is a binary function which returns true if the history h has feature x. We focus on the attributes of a bunsetsu itself and on the features occurring between bunsetsus.

Given a set of features and some training data, the maximum entropy estimation process produces a model in which every feature g_i has associated with it a parameter α_i. This allows us to compute the conditional probability as follows (Berger et al., 1996):

$$P(f|h) = \frac{\prod_i \alpha_i^{g_i(h,f)}}{Z_\lambda(h)} \quad (2)$$

$$Z_\lambda(h) = \sum_f \prod_i \alpha_i^{g_i(h,f)}. \quad (3)$$

The maximum entropy estimation technique guarantees that for every feature g_i, the expected value of g_i according to the M.E. model will equal the empirical expectation of g_i in the training corpus. In other words:

$$\sum_{h,f} \tilde{P}(h,f) \cdot g_i(h,f) = \sum_h \tilde{P}(h) \cdot \sum_f P_{ME}(f|h) \cdot g_i(h,f). \quad (4)$$

Here $\tilde{P}$ is an empirical probability and P_{ME} is the

Table 1: Example of estimating the probabilities of word orders.

"昨日 (yesterday) / 太郎は (Taro) / テニスを (tennis) / した。 (played.)"	$P_{\text{昨日},\text{太郎は}} \times P_{\text{昨日},\text{テニスを}} \times P_{\text{太郎は},\text{テニスを}} = 0.6 \times 0.8 \times 0.7 = 0.336$
"昨日 (yesterday) / テニスを (tennis) / 太郎は (Taro) / した。 (played.)"	$P_{\text{昨日},\text{太郎は}} \times P_{\text{昨日},\text{テニスを}} \times P_{\text{テニスを},\text{太郎は}} = 0.6 \times 0.8 \times 0.3 = 0.144$
"太郎は (Taro) / 昨日 (yesterday) / テニスを (tennis) / した。 (played.)"	$P_{\text{太郎は},\text{昨日}} \times P_{\text{昨日},\text{テニスを}} \times P_{\text{太郎は},\text{テニスを}} = 0.4 \times 0.8 \times 0.7 = 0.224$
"太郎は (Taro) / テニスを (tennis) / 昨日 (yesterday) / した。 (played.)"	$P_{\text{太郎は},\text{昨日}} \times P_{\text{テニスを},\text{昨日}} \times P_{\text{太郎は},\text{テニスを}} = 0.4 \times 0.2 \times 0.7 = 0.056$
"テニスを (tennis) / 昨日 (yesterday) / 太郎は (Taro) / した。 (played.)"	$P_{\text{昨日},\text{太郎は}} \times P_{\text{テニスを},\text{昨日}} \times P_{\text{テニスを},\text{太郎は}} = 0.6 \times 0.2 \times 0.3 = 0.036$
"テニスを (tennis) / 太郎は (Taro) / 昨日 (yesterday) / した。 (played.)"	$P_{\text{太郎は},\text{昨日}} \times P_{\text{テニスを},\text{昨日}} \times P_{\text{テニスを},\text{太郎は}} = 0.4 \times 0.2 \times 0.3 = 0.024$

probability assigned by the M.E. model.

We define a word order model as a model which learns the appropriate order of each pair of modifiers which depend on the same modifiee. This model is derived from Eq. (2) as follows. Assume that there are two bunsetsus B_1 and B_2 which depend on the bunsetsu B and that h is the information derivable from the test corpus. The probability that "B_1 B_2" is the appropriate order is given by the following equation:

$$P_{ME}(1|h) = \frac{\prod_{i=1}^{k} \alpha_{1,i}^{g_i(1,b)}}{\prod_{i=1}^{k} \alpha_{1,i}^{g_i(1,b)} + \prod_{i=1}^{k} \alpha_{0,i}^{g_i(0,b)}}, \quad (5)$$

where $g_i(1 \leq i \leq k)$ is a feature and "1" indicates that the order is appropriate. The terms $\alpha_{1,i}$ and $\alpha_{0,i}$ are estimated from a corpus which is morphologically and syntactically analyzed. When there are three or more bunsetsus that depend on the same modifiee, the probability is estimated as follows: For n bunsetsus B_1, B_2, …, B_n which depend on the bunsetsu B and for the information h derivable from the test corpus, the probability that "B_1 B_2 … B_n" is the appropriate order, or $P(1|h)$, is represented as the probability that every two bunsetsus "B_i B_{i+j} $(1 \leq i \leq n-1, 1 \leq j \leq n-i)$" are the appropriate order, or $P(\{W_{i,i+j} = 1 | 1 \leq i \leq n-1, 1 \leq j \leq n-i\}|h)$. Here "$W_{i,i+j} = 1$" represents that "$B_i$ B_{i+j}" is the appropriate order. Let us assume that every $W_{i,i+j}$ is independent each other. Then $P(1|h)$ is derived as follows:

$$\begin{aligned} P(1|h) &= P(\{W_{i,i+j} = 1 | 1 \leq i \leq n-1, \\ &\qquad\qquad 1 \leq j \leq n-i\}|h) \\ &\approx \prod_{i=1}^{n-1} \prod_{j=1}^{n-i} P(W_{i,i+j} = 1 | h_{i,i+j}) \\ &= \prod_{i=1}^{n-1} \prod_{j=1}^{n-i} P_{ME}(1|h_{i,i+j}), \quad (6) \end{aligned}$$

where $h_{i,i+j}$ is the information derivable when focusing on the bunsetsu B and its modifiers B_i and B_{i+j}.

For example, in the sentence "昨日 (*kinou*, yesterday) / 太郎は (*Taro_wa*, Taro) / テニスを (*tennis_wo*, tennis) / した。 (*sita.*, played.)," where a "/" represents a bunsetsu boundary, there are three bunsetsus that depend on the verb "した (*sita*)." We train a word order model under the assumption that the orders of three pairs of modifiers —"昨日" and "太郎は," "昨日" and "テニスを," and "太郎は" and "テニスを"— are appropriate. We use various kinds of information in and around the target bunsetsus as features. For example, the information or the feature that a noun of time precedes a proper noun is derivable from the order "昨日 (yesterday) / 太郎は (Taro) / した。 (played.)," and the feature that a case followed by a case marker "wa" precedes a case followed by a case marker "wo" is derivable from the order "太郎は (*Taro_wa*, Taro) / テニスを (*tennis_wo*, tennis) / した。 (*sita.*, played.)."

2.2 Word Order Estimation

This section describes the algorithm of estimating the word order by using a trained word order model. The word order estimation is defined as deciding the order of modifiers or bunsetsus which depend on the same modifiee. The input of this task consists of modifiers and information necessary to know whether or not features are found. The output is the order of the modifiers. We assume that lexical selection in each bunsetsu is already done and all dependencies in a sentence are found. The information necessary to know whether or not features are found is morphological, syntactic, semantic, and contextual information, and the locations of bunsetsu boundaries. The features used in our experiments are described in Section 3.

Word order is estimated in the following steps.

Procedures

1. All possible orders of modifiers are found.
2. For each, the probability that it is appropriate is estimated by a word order model, or Eq. (6).
3. The order with the highest probability of being appropriate is selected.

For example, given the sentence "昨日 (*kinou*, yesterday) / 太郎は (*Taro_wa*, Taro) / テニスを (*tennis_wo*, tennis) / した。 (*sita.*, played.)," the modifiers of a verb "した (played)" are three bunsetsus, "昨日 (yesterday)," "太郎は (Taro)," "テニスを (tennis)." Their appropriate order is estimated in the following steps.

1. The probabilities that the orders of the three pairs of modifiers "昨日" and "太郎は," "昨日" and "テニスを," and "太郎は" and "テニスを" are appropriate are estimated. Assume, for example, $P_{\text{昨日},\text{太郎は}}$, $P_{\text{昨日},\text{テニスを}}$, and $P_{\text{太郎は},\text{テニスを}}$ are respectively 0.6, 0.8, and 0.7.
2. As shown in Table 1, probabilities are estimated for all six possible orders. The order "昨日 / 太郎は / テニスを / した。," which has the highest probability, is selected as the most appropriate order.

2.3 Performance Evaluation

The *performance* of a word order model can be evaluated in the following way. First, extract from a test corpus bunsetsus having two or more modifiers. Then, using those bunsetsus and their modifiers as

Table 2: Example of modifiers extracted from a corpus.

Data				Modifiers (Bunsetsu number)
Bunsetsu number	Bunsetsu number of modifier	Label	Strings in a bunsetsu	Modifiers whose modifiee is the bunsetsu in the left column.
0	1	P	太郎と (*Taro_to*, Taro and)	
1	5		花子は (*Hanako_to*, Hanako)	太郎と (0)
2	3		テニスの (*tennis_no*, tennis)	
3	4		試合に (*siai_ni*, tournament)	テニスの (2)
4	5	P	出て、 (*dete*,, participate,)	太郎と (0) 花子は (1) 試合に (3)
5			優勝した。 (*yusyo_sita.*, won.)	太郎と (0) 花子は (1) 出て (4)

input, estimate the orders of the modifiers as described in Section 2.2. The percentage of the modifiees whose modifiers' word order agrees with that in the original text then gives what we call the *agreement rate*. It is a measure of how close the word order estimated by the model is to the actual word order in the training corpus.

We use the following two measurements to calculate the agreement rate.

Pair of modifiers The first measurement is the percentage of the pairs of modifiers whose word order agrees with that in the test corpus. For example, given the sentence in a test corpus "昨日 (*kinou*, yesterday) / 太郎は (*Taro_wa*, Taro) / テニスを (*tennis_wo*, tennis) / した。 (*sita.*, played.)," if the word order estimated by the model is "昨日 (yesterday) / テニスを (tennis) / 太郎は (Taro) / した。 (played.)," then the orders of the pairs of modifiers in the original sentence are "昨日 / 太郎は," "昨日 / テニスを," and "太郎は / テニスを," and those in the estimated word order are "昨日 / テニスを," "昨日 / 太郎は," and "テニスを / 太郎は." The agreement rate is 67% (2/3) because two of the three orders are the same as those in the original sentence.

Complete agreement The second measurement is the percentage of the modifiees whose modifiers' word order agrees with that in the test corpus.

3 Experiments and Discussion

In our experiment, we used the Kyoto University text corpus (Version 2) (Kurohashi and Nagao, 1997), a tagged corpus of the Mainichi newspaper. For training, we used 17,562 sentences from newspaper articles appearing in 1995, from January 1st to January 8th and from January 10th to June 9th. For testing, we used 2,394 sentences from articles appearing on January 9th and from June 10th to June 30th.

3.1 Definition of Word Order in a Corpus

In the Kyoto University corpus, each bunsetsu has only one modifiee. When a bunsetsu B_m depends on a bunsetsu B_d and there is a bunsetsu B_p that depends on and is coordinate with B_d, B_p has not only the information that its modifiee is B_d but also a label indicating a coordination or the information that it is coordinate with B_d. This information indirectly shows that the bunsetsu B_m can depend on both B_p and B_d. In this case, we consider B_m a modifier of both B_p and B_d.

Under this condition, modifiers of a bunsetsu B are identified in the following steps.

1. Bunsetsus that depend on a bunsetsu B are classified as modifiers of B.
2. When B has a label indicating a coordination, bunsetsus that are to the left of B and depend on the same modifiee as B are classified as modifiers of B.
3. Bunsetsus that depend on a modifier of B and have a label indicating a coordination are classified as modifiers of B. The third step is repeated.

When the above procedure is completed, all bunsetsus that coordinate with each other are identified as modifiers which depend on the same modifiee. For example, from the data listed on the left side of Table 2, the modifiers listed in the right-hand column are identified for each bunsetsu. "太郎と (*Taro_to*, Taro and)," "花子は (*Hanako_to*, Hanako)," "出て、 (*dete*,, participate,)" are all identified as modifiers which depend on the same modifiee "優勝した。 (*yusyo_sita.*, won.)."

3.2 Experimental Results

The features used in our experiment are listed in Tables 3 and 4. Each feature consists of a type and a value. The features consist basically of some attributes of the bunsetsu itself, and syntactic and contextual information. We call the features listed in Tables 3 'basic features.' We selected them manually so that they reflect the basic conditions governing word order that were summarized by Saeki (Saeki, 1998). The features in Table 4 are combinations of basic features ('combined features') and were also selected manually. They are represented by the name of the target bunsetsu plus the feature type of the basic features. The total number of features was about 190,000, and 51,590 of them were observed in the training corpus three or more times. These were the ones we used in our experiment.

The following terms are used in these tables:

Mdfr1, Mdfr2, Mdfe: The word order model described in Section 2.1 estimates the probability that modifiers are in the appropriate order as the product of the probabilities of all pairs of modifiers. When estimating the probability for each pair of modifiers, the model assumes that the two modifiers are in the appropriate order. Here we call the left modifier Mdfr1, the right modifier Mdfr2, and their modifiee Mdfe.

Head: the rightmost word in a bunsetsu other than those whose major part-of-speech[1] category is

[1] Part-of-speech categories follow those of JUMAN (Kurohashi and Nagao, 1998).

Table 3: Basic features.

Basic features				Accuracy without each category	
Category	Target bunsetsus	Feature type	Feature values (Number of type)	Pair of modifiers	Complete agreement
1	Mdfr1, Mdfr2, Mdfe	Head-Lex	(5,066)	86.65% (−0.79%)	73.87% (−1.54%)

2	Mdfr1, Mdfr2, Mdfe	Head-POS(Major)
		Head-POS(Minor)
			動詞 (verb), 形容詞 (adjective), 名詞 (noun), . . . (11)
			普通名詞 (common noun), 数詞 (quantifier), . . . (24)
				87.07%
				(−0.37%)
					75.03%
					(−0.38%)

3	Mdfr1, Mdfr2, Mdfe	Head-Inf(Major)
		Head-Inf(Minor)
			母音動詞 (vowel verb), . . . (30)
			語幹 (stem), 基本形 (fundamental form), . . . (60)
				87.39%
				(−0.05%)
					75.20%
					(−0.21%)

4	Mdfr1, Mdfr2, Mdfe	Head-SemFeat(110)
		Head-SemFeat(111)
		. . .
		Head-SemFeat(433) (Total : 90)
			True (1)
			True (1)
			
			True (1)
				87.21%
				(−0.23%)
					75.20%
					(−0.21%)

5	Mdfr1, Mdfr2, Mdfe	Type(String)
		Type(Major)
		Type(Minor)
			こそ, こと, そして, だけ, と, に, も, . . . (73)
			助詞 (post-positional particle), . . . (43)
			格助詞 (case marker), 命令形 (imperative form) . . . (102)
				84.78%
				(−2.66%)
					70.03%
					(−5.38%)

6	Mdfr1, Mdfr2, Mdfe	JOSHI1(String)
		JOSHI1(Minor)
		JOSHI2(String)
		JOSHI2(Minor)
			から, まで, のみ, へ, ねえ, . . . (63)
			[nil], 格助詞 (case marker), . . . (5)
			けど, まま, や, よ, か, . . . (63)
			格助詞 (case marker), . . . (4)
				87.32%
				(−0.12%)
					75.14%
					(−0.27%)

7	Mdfr1, Mdfr2, Mdfe	Period	[nil], [exist] (2)	87.39% (−0.05%)	75.54% (+0.13%)
8	Mdfr1, Mdfr2	NumberOfMdfrs	A(0), B(1), C(2), D(3 or more) (4)	87.14% (−0.30%)	74.86% (−0.55%)
	Mdfe	NumberOfMdfrs	A(2), B(3), C(4 or more) (3)	87.40% (−0.04%)	75.35% (−0.06%)
9	Mdfr1, Mdfr2, Mdfe	Coordination	P(Coordinate), A(Apposition), D(otherwise) (3)	86.26% (−1.18%)	73.61% (−1.80%)
10	Mdfr1, Mdfr2	Mdfr1-MdfrType-IDto-Mdfr2-Type Mdfr2-MdfrType-IDto-Mdfr1-Type Mdfr1-MdfrType-IDto-Mdfr2-MdfrType	True, False (2) True, False (2) True, False (2)	87.34% (−0.10%)	75.09% (−0.32%)
11	Mdfr1, Mdfr2, Mdfe	Repetition-Head-Lex Repetition-Mdfr-Head-Lex	[nil], [exist] (2) [nil], [exist] (2)	87.31% (−0.13%)	75.14% (−0.27%)
12	Mdfr1, Mdfr2	ReferencePronoun ReferencePronoun(String)	[nil], [exist] (2) この, これ, こんな, そこ, その, それ, . . . (42)	87.27% (−0.17%)	75.12% (−0.29%)

"特殊 (special marks)," "助詞 (post-positional particles)," or "接尾辞 (suffixes)."

Head-Lex: the fundamental form (uninflected form) of the head word. Only words with a frequency of five or more are used.

Head-Inf: the inflection type of a head.

SemFeat: We use the upper third layers of *bunrui goihyou* (NLRI(National Language Research Institute), 1964) as semantic features. Bunrui goihyou is a Japanese thesaurus that has a tree structure and consists of seven layers. The tree has words in its leaves, and each word has a figure indicating its category number. For example, the figure in parenthesis of a feature "Head-SemFeat(110)" in Table 3 shows the upper three digits of the category number of the head word or the ancestor node of the head word in the third layer in the tree.

Type: the rightmost word other than those whose major part-of-speech category is "特殊 (special marks)." If the major category of the word is neither "助詞 (post-positional particles)" nor "接尾辞 (suffixes)," and the word is inflectable,[2] then the type is represented by the inflection type.

JOSHI1, JOSHI2: JOSHI1 is the rightmost post-positional particle in the bunsetsu. And if there are two or more post-positional particles in the bunsetsu, JOSHI2 is the second-rightmost post-positional particle.

NumberOfMdfrs: number of modifiers.

Mdfr1-MdfrType, Mdfr2-MdfrType: Types of the modifiers of Mdfr1 and Mdfr2.

X-IDto-Y: X is identical to Y.

Repetition-Head-Lex: a repetition word appearing in a preceding sentence.

ReferencePronoun: a reference pronoun appearing in the target bunsetsu or in its modifiers.

Categories 1 to 6 in Table 3 represent attributes in a bunsetsu, categories 7 to 10 represent syntactic information, and categories 11 and 12 represent contextual information.

The results of our experiment are listed in Table 5. The first line shows the agreement rate when we estimated word order for 5,278 bunsetsus that have two or more modifiers and were extracted from 2,394 sentences appearing on January 9th and from June 10th to June 30th. We used bunsetsu boundary information and syntactic and contextual information which were derivable from the test corpus and related to the input bunsetsus. As syntactic information we used dependency information, coordinate structure, and information on whether the target bunsetsu is at the end of a sentence. As contextual information we used the preceding sentence. The values in the row labeled Baseline1 in Table 5 are the agreement rates obtained when every order of all pairs of modifiers was selected randomly. And values in the Baseline2 row are the agreement rates obtained when we used the following equation instead of Eq. (5):

$$P_{ME}(1|h) \;=\; \frac{freq(w_{12})}{freq(w_{12}) + freq(w_{21})}. \qquad (7)$$

[2] The inflection types follow those of JUMAN.

Table 4: Combined features.

Combined features	Accuracy without the feature	
	Pair of modifiers	Complete agreement
Twin features	87.23% (−0.21%)	74.65% (−0.76%)
(Mdfr1-Type, Mdfr2-Type), (Mdfr1-Type, Mdfe-Head-Lex), (Mdfr1-Type, Mdfe-Head-POS), (Mdfr1-Type, Mdfr1-Coordination), (Mdfr1-Type, Mdfr2-MdfrType-IDto-Mdfr1-Type), (Mdfr2-Type, Mdfe-Head-Lex), (Mdfr2-Type, Mdfe-Head-POS), (Mdfr2-Type, Mdfr2-Coordination), (Mdfr2-Type, Mdfr1-MdfrType-IDto-Mdfr2-Type), (Mdfr1-Head-Lex, Mdfe-Period), (Mdfr1-Head-POS, Mdfe-Period), (Mdfr1-Head-POS, Mdfr1-Repetition-Head-Lex), (Mdfr2-Head-Lex, Mdfe-Period), (Mdfr2-Head-POS, Mdfe-Period), (Mdfr2-Head-POS, Mdfr2-Repetition-Head-Lex)		
Triplet features	87.22% (−0.22%)	74.86% (−0.55%)
(Mdfr1-Type, Mdfr2-Type, Mdfe-Head-Lex), (Mdfr1-Type, Mdfr2-Type, Mdfe-Head-POS), (Mdfr1-Type, Mdfr1-Coordination, Mdfe-Type), (Mdfr2-Type, Mdfr2-Coordination, Mdfe-Type), (Mdfr1-JOSHI1, Mdfr1-JOSHI2, Mdfe-Head-Lex), (Mdfr1-JOSHI1, Mdfr1-JOSHI2, Mdfe-Head-POS), (Mdfr2-JOSHI1, Mdfr2-JOSHI2, Mdfe-Head-Lex), (Mdfr2-JOSHI1, Mdfr2-JOSHI2, Mdfe-Head-POS)		
All of above combined features	85.79% (−1.65%)	71.67% (−3.74%)

Table 5: Results of agreement rates.

	Agreement rate	
	Pair of modifiers	Complete agreement
Our method	87.44%(12,361/14,137)	75.41% (3,980/5,278)
Baseline1	48.96% (6,921/14,137)	33.10% (1,747/5,278)
Baseline2	49.20% (6,956/14,137)	33.84% (1,786/5,278)

Here we assume that B_1 and B_2 are modifiers, their modifiee is B, the word types of B_1 and B_2 are respectively w_1 and w_2. The values $freq(w_{12})$ and $freq(w_{21})$ then respectively represent the frequencies with which w_1 and w_2 appeared in the order "w_1, w_2, and w" and "w_2, w_1, and w" in Mainichi newspaper articles from 1991 to 1997. [3] Equation (7) means that given the sentence "太郎は (*Taro_wa*) / テニスを (*tennis_wo*) / した。 (*sita.*)," one of two possibilities, "は (*wa*) / を (*wo*) / した。 (*sita.*)" and "を (*wo*) / は (*wa*) / した。 (*sita.*)," which has the higher frequency, is selected.

3.3 Features and Agreement Rate

This section describes how much each feature set contributes to improving the agreement rate.

The values listed in the rightmost columns in Tables 3 and 4 shows the performance of the word order estimation without each feature set. The values in parentheses are the percentage of improvement or degradation to the formal experiment. In the experiments, when a basic feature was deleted, the combined features that included the basic feature were also deleted. The most useful feature is the type of bunsetsu, which basically signifies the case marker or inflection type. This result is close to our expectations.

We selected features that, according to linguistic studies, as much as possible reflect the basic conditions governing word order. The rightmost column in Tables 3 and 4 shows the extent to which each condition contributes to improving the agreement rate. However, each category of features might be rougher than that which is linguistically interesting. For example, all case markers such as "wa" and "wo" were classified into the same category, and were deleted together in the experiment when single categories were removed. An experiment that considers each of these markers separately would help us verify the importance of these markers separately. If we find new features in future linguistic research on word order, the experiments lacking each feature separately would help us verify their importance in the same manner.

3.4 Training Corpus and Agreement Rate

The agreement rates for the training corpus and the test corpus are shown in Figure 1 as a function of the amount of training data (number of sentences). The agreement rates in the "pair of modifiers" and "Complete agreement" measurements were respectively 82.54% and 68.40%. These values were obtained with very small training sets (250 sentences). These rates are considerably higher than those of the baselines, indicating that word order in Japanese can be acquired from newspaper articles even with a small training set.

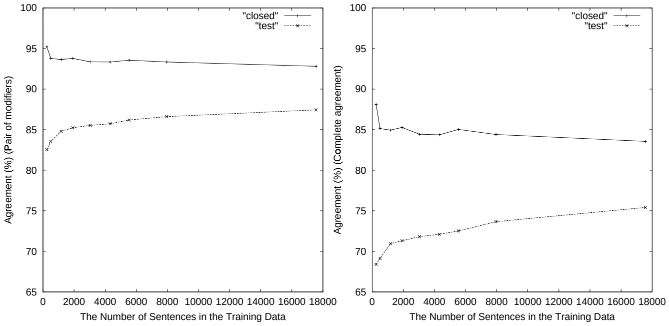

Figure 1: Relationship between the amount of training data and the agreement rate.

With 17,562 training sentences, the agreement rate in the "Complete agreement" measurement was 75.41%. We randomly selected and analyzed 100 modifiees from 1,298 modifiees whose modifiers' word order did not agree with those in the original text. We found that 48 of them were in a natural order and 52 of them were in an unnatural order. The former result shows that the word order was relatively free and several orders were acceptable. The latter result shows that the word order acquisition was not sufficient. To complete the acquisition we need more training corpora and features which take into account different information than that in Tables 3 and 4. We found many idiomatic expres-

[3] When w_1 and w_2 were the same word, we used the head words in B_1 and B_2 as w_1 and w_2. When one of $freq(w_{12})$ and $freq(w_{21})$ was zero and the other was five or more, we used the frequencies when they appeared in the order "w_1 w_2" and "w_2 w_1," respectively, instead of $freq(w_{12})$ and $freq(w_{21})$. When both $freq(w_{12})$ and $freq(w_{21})$ were zero, we instead used random figures between 0 and 1.

sions in the unnatural word order results, such as “法治国家が (*houchi-kokka_ga*, a country under the rule of law) / 聞いて (*kiite*, to listen) / あきれる (*akireru*, to disgust),” “創案したのが (*souan-sita-no_ga*, origination) / そもそもの (*somosomo-no*, at all) / 始まり (*hajimari*, the beginning),” and “味に (*aji_ni*, taste) / 精魂 (*seikon*, one’s heart and soul) / 込める (*komeru*, to put something into something).” We think that the appropriate word order for these idiomatic expressions could be acquired if we had more training data. We also found several coordinate structures in the unnatural word order results, suggesting that we should survey linguistic studies on coordinate structures and try to find efficient features for acquiring word order from coordinate structures.

We did not use the results of semantic and contextual analyses as input because corpora with semantic and contextual tags were not available. If such corpora were available, we could more efficiently use features dealing with semantic features, reference pronouns, and repetition words. We plan to make corpora with semantic and contextual tags and use these tags as input.

3.5 Acquisition from a Raw Corpus

In this section, we show that a raw corpus instead of a tagged corpus can be used to train the model, if it is first analyzed by a parser. We used the morphological analyzer JUMAN and a parser KNP (Kurohashi, 1998) which is based on a dependency grammar, in order to extract information from a raw corpus for detecting whether or not each feature is found. The accuracy of JUMAN for detecting morphological boundaries and part-of-speech tags is about 98%, and the parser’s dependency accuracy is about 90%. These results were obtained from analyzing Mainichi newspaper articles.

We used 217,562 sentences for training. When these sentences were all extracted from a raw corpus, the agreement rate was 87.64% for “pair of modifiers” and was 75.77% for “Complete agreement.” When the 217,562 training sentences were sentences from the tagged corpus (17,562 sentences) used in our formal experiment and from a raw corpus, the agreement rate for “pair of modifiers” was 87.66% and for “Complete agreement” was 75.88%. These rates were about 0.5% higher than those obtained when we used only sentences from a tagged corpus. Thus, we can acquire word order by adding information from a raw corpus even if we do not have a large tagged corpus. The results also indicate that the parser accuracy is not so significant for word order acquisition and that an accuracy of about 90% is sufficient.

4 Conclusion

This paper described a method of acquiring word order from corpora. We defined word order as the order of modifiers which depend on the same modifiee. The method uses a model which estimates the likelihood of the appropriate word order. The model automatically discovers what the tendency of the word order in Japanese is by using various kinds of information in and around the target bunsetsus plus syntactic and contextual information. The contribution rate of each piece of information in deciding word order is efficiently learned by a model implemented within an M.E. framework. Comparing results of experiments controlling for each piece of information, we found that the type of information having the greatest influence was the case marker or inflection type in a bunsetsu. Analyzing the relationship between the amount of training data and the agreement rate, we found that word order could be acquired even with a small set of training data. We also found that a raw corpus as well as a tagged corpus can be used to train the model, if it is first analyzed by a parser. The agreement rate was 75.41% for the Kyoto University corpus. We analyzed the modifiees whose modifiers’ word order did not agree with that in the original text, and found that 48% of them were in a natural order. This shows that, in many cases, word order in Japanese is relatively free and several orders are acceptable.

The text we used were newspaper articles, which tend to have a standard word order, but we think that word orders tend to differ between different styles of writing. We would therefore like to carry out experiments with other types of texts, such as novels, having styles different from that of newspapers.

It has been difficult to evaluate the results of text generation objectively because there have been no good standards for evaluation. By using the standard we describe in this paper, however, we can evaluate results objectively, at least for word order estimation in text generation.

We expect that our model can be used for several applications as well as linguistic verification, such as text refinement support and text generation in machine translation.

References

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71.

Sadao Kurohashi and Makoto Nagao. 1997. Kyoto University Text Corpus Project. In *Proceedings of The Third Annual Meeting of The Association for Natural Language Processing*, pages 115–118. (in Japanese).

Sadao Kurohashi and Makoto Nagao, 1998. *Japanese Morphological Analysis System JUMAN Version 3.6.* Department of Informatics, Kyoto University.

Sadao Kurohashi, 1998. *Japanese Dependency/Case Structure Analyzer KNP Version 2.0b6.* Department of Informatics, Kyoto University.

Hiroshi Maruyama. 1994. Experiments on Word-Order Recovery Using N-Gram Models. In *The 49th Annual Convention IPS Japan.* (in Japanese).

NLRI(National Language Research Institute). 1964. *Word List by Semantic Principles.* Syuei Syuppan. (in Japanese).

Tetsuo Saeki. 1998. *Yousetsu nihongo no gojun (Survey: Word Order in Japanese).* Kuroshio Syuppan. (in Japanese).

James Shaw and Vasileios Hatzivassiloglou. 1999. Ordering Among Premodifiers. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 135–143.

Takenobu Tokunaga and Hozumi Tanaka. 1991. On Estimating Japanese Word Order Based on Valency Information. *Keiryo Kokugogaku (Mathematical Linguistics)*, 18(2):53–65. (in Japanese).