Word Order Acquisition from Corpora

Kiyotaka Uchimoto¹, Masaki Murata¹, Qing Ma¹,
Satoshi Sekine⁴, and Hitoshi Isahara⁴

¹Communications 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

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 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.

Componental conditions

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

• A bunsetsu having wide dependency tends to precede a bunsetsu having narrow 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
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 re-formulate 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

$$g(h,f) = \begin{cases} 1 : & \text{if has}(h,x) = \text{true}, \\ & x = \text{"Mdfrl - Head-POS(Major) : verb"} \\ & \text{& } f = 1 \\ 0 : & \text{otherwise.} \end{cases}$$

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 $\alpha_i$. This allows us to compute the conditional probability as follows (Berger et al., 1996):

$$P(f|h) = \frac{\prod_i \alpha_i^{g_i(h,f)}}{Z_h(h)}$$

$$Z_h(h) = \sum_f \prod_i \alpha_i^{g_i(h,f)}.$$ 

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:

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

Here $\tilde{P}$ is an empirical probability and $P_{ME}$ is the
We define a word order model as a model which learns the appropriate order of each pair of modifiers which depend on the same modifier. 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:
\[
P_{MR}(1|h) = \frac{1}{k} \prod_{i=1}^{k} \alpha_{i,j}(h, i, j) + \prod_{i=1}^{k} \alpha_{0,i}(0, i), \tag{5}
\]
where $\alpha_i(j)$ is a feature and "1" indicates that the order is appropriate. The terms $\alpha_{i,j}$ 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 modifier, the probability is estimated as follows: For $n$ bunsetsus $B_1, B_2, \ldots, B_n$ which depend on the bunsetsu $B$ and for the information $h$ derivable from the test corpus, the probability that "$B_1 B_2 \ldots B_n$" 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 $P(W_i, h, i, j) = 1$ for all $i$, is estimated by a word order model, or Eq. (6).

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 (Taro) / テニスを (tennis, play) / した (played)." the modifiers of a verb "した (played)" are three bunsetsus, "昨日 (kinou, yesterday), "太郎 (Taro)," テニスを (tennis, play)." Their appropriate order is estimated in the following steps.

1. The probabilities that the orders of the three pairs of modifiers "昨曰" and "太郎" are "テニスを," and "太郎" and "テニスを" are appropriate are estimated. Assume, for example, $P_{MR}(1\text{, 昨日, 太郎, テニスを})$ and $P_{MR}(1\text{, 太郎, テニスを, 昨日})$ 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 "昨曰 / 太郎 / テニスを / した, (played)," 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
input, estimate the orders of the modifiers as described in Section 2.2. The percentage of the modifiers whose 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 "太郎 (Taro) / は (be) / 今日 (today) / テニスを (tennis to) / した (did) ." if the word order estimated by the model is "Taro / は (be) / 今日 (today) / テニスを (tennis to) / した (did) ." then the orders of the pairs of modifiers in the original sentence are "Taro / は (be)" and "今日 / テニスを (tennis to)" and the estimated word order are "Taro / テニスを (tennis to) / は (be) / 今日 (today)" and "今日 / テニスを (tennis to) / は (be) / Taro ." 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 modifiers whose 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 modifier. 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 modifier 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 modifier 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 classified as modifiers which depend on the same modifier. 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 (Taro) / は (be) / 今日 (today) / テニスを (tennis to) / した (did) ." "太郎 (Taro) / テニスを (tennis to) / した (did) ." "いい (good) / と (and) / は (be) / カード (card) / した (did) ." "は (be) / と (and) / カード (card) / した (did) ." "いい (good) / と (and) / は (be) / カード (card) / した (did) ."

### 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 modifier **Mdfe**.

- **Head**: the rightmost word in a bunsetsu other than those whose major part-of-speech\(^1\) category is

| Bunsetsu number | Bunsetsu number of modifier | Label | Strings in a bunsetsu | Modifiers whose modifier is the bunsetsu in the left column |
|----------------|-----------------------------|-------|----------------------|------------------------------------------------------------|
| 0              | 1                           | P     | 太郎と (Taro.to, Taro and) | 太郎と (Taro) |
| 1              | 5                           |       | は (be)              | は (be)          |
| 2              | 3                           |       | テニスを (tennis to)  | テニスを (tennis to) |
| 3              | 4                           |       | した (did)           | した (did)       |
| 4              | 5                           | P     | と (and)             | と (and)          |

---

1Part-of-speech categories follow those of JUMAN (Kurohashi and Nagao, 1998).
Table 3: Basic features.

| Category | Target bunsetsus | Feature type | Feature values (Number of type) | Accuracy without each category | Pair of modifiers agreement | Complete agreement |
|----------|------------------|--------------|--------------------------------|-------------------------------|-----------------------------|-------------------|
| 1        | Mdfr1, Mdfr2     | Head-Lex     | [0,060]                        | 86.65% (0.70%)                | 75.87% (-1.54%)            |
| 2        | Mdfr1, Mdfr2     | Head-POS(Major) | [3,063]                        | 87.63% (0.77%)                | 75.32% (-1.38%)            |
| 3        | Mdfr1, Mdfr2     | Head-Inf(Major) |                          | 87.39% (0.70%)                | 75.20% (-0.21%)            |
| 4        | Mdfr1, Mdfr2     | Head-Inf(Minor) |                          | 87.21% (0.70%)                | 75.20% (-0.21%)            |

- **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 bunsetsus (NRLI(National Language Research Institute), 1964) as semantic features. Bunsetsu 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),” “助詞 (post-positional particles),” or “助動詞 (suffixes).”
- **ReferencePronoun**: a reference pronoun appearing in a preceding sentence.
- **Mdf1-MdfrType, Mdf2-MdfrType**: Types of the modifiers of Mdfr1 and Mdfr2.
- **Coordination**

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 derived from the test corpus and related to the input bunsetsus. As contextual 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 Baseline 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):

\[ P_{M}(h|l) = \frac{freq(w_{12})}{freq(w_{12}) + freq(w_{21})} \]
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 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 "all of above combined features" are shown in Table 5.

Here we assume that $B_1$ and $B_2$ are modifiers, their modifyee is $B$, the word types of $B_1$ and $B_2$ are respectively $w_1$ and $w_2$. The values $freq(w_1)$ and $freq(w_2)$ then respectively represent the frequencies with which $w_1$ and $w_2$ appeared in the order "$w_1$, $w_2$", and "$w_2$, $w_1$", and "$w_1$, $w_2$", and "$w_2$, $w_1$" in Mainichi newspaper articles from 1991 to 1997. Equation (7) means that given the sentence "tare wa (Taro_wa) / テニス を (tennis_wo) にした。", one of two possibilities, "wa (wa) / を (wo) にした。" and "を (wo) / にした。 (sita).", which has the higher frequency, is selected.

### Table 4: Combined features.

| Combined features | Accuracy without the feature | Pair of modifiers | Complete agreement |
|-------------------|-----------------------------|------------------|--------------------|
| Twin features     |                             | 87.23% (-0.33%)  | 74.86% (-0.55%)    |
| Triple features   |                             | 87.22% (-0.32%)  | 74.86% (-0.55%)    |
| All of above combined features | 86.79% (-1.65%) | 71.67% (-3.74%) |

### Table 5: Results of agreement rates.

| Agreement rate | Pair of modifiers | Complete agreement |
|----------------|-------------------|-------------------|
| Our method     | 87.44% (6,921/14,137) | 75.41% (3,888/5,278) |
| Baseline1      | 49.20% (6,921/14,137)  | 35.10% (7,747/5,278)  |
| Baseline2      | 49.20% (6,921/14,137)  | 35.10% (7,747/5,278)  |

The values listed in the rightmost columns 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.

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

"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.

With 17,562 training sentences, the agreement rate in the "Complete agreement" measurement was 75.41%. We randomly selected and analyzed 100 modifyees from 1,298 modifyees 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-
sions in the unnatural word order results, such as "法
治国家 (honchi-kokka-ga, a country under the rule
of law) / 伺って (kitte, to listen) / してされる (akseru,
to disgust)." "創造したのが (souzan-situ-no-ga, origi-
nation) / それぞれの (somosomo-no, at all) / 始まり
(hajimari, the beginning)," and "味に (aji, ni, taste) /
精魂 (seikon, one's heart and soul) / 込む (konaru,
to put something into something)." We think that
the appropriate word order for these idiomatic ex-
pressions 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 struc-
tures and try to find efficient features for acquiring
word order from coordinate structures.
We did not use the results of semantic and con-
textual analyses as input because corpora with se-
matic 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
use features dealing with semantic features, reference
pronouns, and repetition words. We plan to make
use of 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 morpho-
logical 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 mor-
phological 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 for-
mal experiment and from a raw corpus, the agree-
ment 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 ac-
curacy 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 or-
der from corpora. We defined word order as the order
of modifiers which depend on the same modifie.
The method uses a model which estimates the likelihood
of the appropriate word order. The model automati-
cally discovers what the tendency of the word order
in Japanese is by using various kinds of information
in and around the target bunsetsu 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 experi-
ments controlling for each piece of information, we
found that the type of information having the great-
est 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 modifiers 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 lack 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 newspa-
pers.

It has been difficult to evaluate the results of text
generation objectively because there have been no
good standards for evaluation. By using the stan-
dard we describe in this paper, however, we can eval-
uate results objectively, at least for word order esti-
mation 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 ma-
chine translation.

References
Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della
Pietra. 1996. A Maximum Entropy Approach to Natural Lan-
guage 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 Process-
ing, pages 115-118. (in Japanese).
Sadao Kurohashi and Makoto Nagao. 1998. Japanese Mor-
phological Analysis System JUMAN Version 3.6. Department
of Informatics, Kyoto University.
Sadao Kurohashi. 1998. Japanese Dependency/Case Structure
Analyzer KNP Version 3.0.6. Department of Informatics, Ky-
oto University.
Hiroshi Maruyama. 1994. Experiments on Word Order Recovery
Using N-Gram Models. In The 14th 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 gojin (Survey: Word
Order in Japanese). Kurosahi Syuppan. (in Japanese).
James Shaw and Vasileios Hatzivassiloglou. 1999. Ordering
Among Premodifiers. In Proceedings of the 37th Annual Meet-
ing of the Association for Computational Linguistics (ACL),
pages 135-143.
Takeshio Toshimana and Hozumi Tamats. 1991. On Estima-
ing Japanese Word Order Based on Valency Information.
Keigo Kokugogaku (Mathematical Linguistics), 18(2):53-65.
(in Japanese).