An Empirical Evaluation on Statistical Parsing of Japanese Sentences using Lexical Association Statistics

SHIRAI Kiyoaki INUI Kentaro TOKUNAGA Takenobu TANAKA Hozumi

Department of Computer Science, Graduate School of Information Science and Engineering,
Tokyo Institute of Technology

Abstract

We are proposing a new framework of statistical language modeling which integrates lexical association statistics with syntactic preference, while maintaining the modularity of those different statistics types, facilitating both training of the model and analysis of its behavior. In this paper, we report the result of an empirical evaluation of our model, where the model is applied to disambiguation of dependency structures of Japanese sentences. We also discussed the room remained for further improvement based on our error analysis.

1 Introduction

In the statistical parsing literature, it has already been established that statistics of lexical association have real potential for improvement of disambiguation performance. The question is how lexical association statistics should be incorporated into the overall statistical parsing framework. In exploring this issue, we consider the following four basic requirements:

- *Integration of different types of statistics*:
  Lexical association statistics should be integrated with other types of statistics that are also expected to be effective in statistical parsing, such as short-term POS n-gram statistics and long-term structural preferences over parse trees.

- *Modularity of statistics types*:
  The total score of a parse derivation should be decomposable into factors derived from different types of statistics, which would facilitate analysis of a model's behavior in terms of each statistics type.

- *Probabilistically well-founded semantics*:
  The language model used in a statistical parser should have probabilistically well-founded semantics, which would also facilitate the analysis of the model's behavior.

- *Trainability*:
  Since incorporation of lexical association statistics would make the model prohibitively complex, the model's complexity should be flexibly controllable depending on the amount of available training data.

However, it seems to be the case that no existing framework of language modeling [2, 4, 12, 13, 14, 17, 18] satisfies these basic requirements simultaneously[1]. In this context, we newly designed a framework of statistical language modeling taking all of the above four requirements into account [8, 9]. This paper reports on the results of our preliminary experiment where our framework was applied to structural disambiguation of Japanese sentences.

In what follows, we first briefly review our framework (Section 2). We next describe the setting of our experiment, including a brief introduction of Japanese dependency structures, the data sets, the baseline of the performance, etc. (Section 3). We then describe the results of the experiment, which was designed to assess the impact of the the incorporation of lexical association statistics (Section 4). We finally discuss the current problems revealed through our error analysis, suggesting some possible solutions (Section 5).

2 Overview of our framework

As with the most statistical parsing frameworks, given an input string A, we rank its parse derivations according to the joint distribution $P(R, W)$, where W is a word sequence candidate for A, and R is a parse derivation candidate for W whose terminal symbols constitute a POS tag sequence L (see Figure 1[2]). We first decompose $P(R, W)$

[1] For further discussion, see [8]. This is also the case with recent works such as [3] and [5] due to the lack of modularity of statistical types.

[2] Although syntactic structure R is represented as a dependency structure in this figure, our framework

into two submodels, the syntactic model $P(R)$ and the lexical model $P(W|R)$:

$$P(R, W) = P(R) \cdot P(W|R) \qquad (1)$$

The syntactic model, which is lexically insensitive, reflects both POS n-gram statistics and structural preference, whereas the lexical model reflects lexical association statistics. This division of labor allows for distinct modularity between the syntactic-based statistics and lexically sensitive statistics, while maintaining the probabilistically well-foundedness of the overall model.

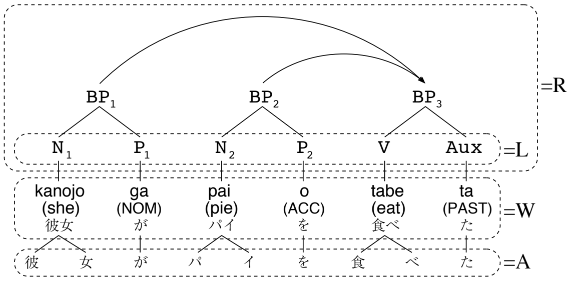

Figure 1: A parse derivation for an input string "彼女がパイを食べた (She ate a pie)"

2.1 The syntactic model

The syntactic model $P(R)$ can be estimated using a wide range of existing syntactic-based language modeling frameworks, from simple PCFG models to more context-sensitive models including those proposed in [2, 13, 19]. Among these, we, at present, use probabilistic GLR (PGLR) language modeling, which is given by incorporating probabilistic distributions into the GLR parsing framework [10, 21]. The advantages of PGLR modeling are (a) PGLR models are mildly context-sensitive, compared with PCFG models, and (b) PGLR models inherently capture both structural preferences and POS bigram statistics, which meets our integration requirement. For further discussion, see [10].

2.2 The lexical model

The lexical model $P(W|R)$ is the product of the probability of each lexical derivation $l_i \rightarrow w_i$, where $l_i \in L$ $(L \subset R)$ is the POS tag of $w_i \in W$:

$$P(W|R) = \prod_i P(w_i|R, w_1, \ldots, w_{i-1}) \qquad (2)$$

The key idea for estimating each factor $P(w_i|R, w_1, \ldots, w_{i-1})$ (a lexical derivation probability) is in assuming that each lexical derivation depends only on a certain small part of its whole context. We first assume that syntactic structure R in $P(w_i|R, w_1, \ldots, w_{i-1})$ can always be reduced to l_i $(\in R)$, which allows us to deal with the lexical model separately from the syntactic model. The question then is which subset C of $\{w_1, \ldots, w_{i-1}\}$ has the strongest influence on the derivation $l_i \rightarrow w_i$. We refer to a member of such a subset C as a *lexical context* of the derivation $l_i \rightarrow w_i$.

Let us illustrate this through the previous example shown in Figure 1. Suppose that the derivation order for W is head-driven, as given below, to guarantee that, for each of the words subordinated by a head word, the context of the derivation of that subordinated word always includes that head word.

ta (PAST) → *tabe* (eat) → *ga* (NOM) → *o* (ACC) → *kanojo* (she) → *pai* (pie)

First, for each lexical item that we don't consider any lexical association, we estimate the probability of its derivation as follows.

$$P(ta|R) \approx P(ta|Aux) \qquad (3)$$

$$P(tabe|R, ta) \approx P(tabe|V) \qquad (4)$$

Second, we estimate the probability of deriving each slot-marker, e.g. "*ga* (NOM)" and "*o* (ACC)", by considering not only the dependency between the head word and each of its slot-markers, but also the dependency between slot-markers subordinated by the same head:

$$P(ga|R, tabe, ta) \approx P(ga|P_1[\mathrm{h}(tabe, [P_1, P_2])]) \qquad (5)$$

$$P(o|R, ga, tabe, ta) \approx P(o|P_2[\mathrm{h}(tabe, [P_1\!:\!ga, P_2])]) \qquad (6)$$

where $\mathrm{h}(h, [s_1, \ldots, s_n])$ is a lexical context denoting a head word h that subordinates the set of slots $s_1, \ldots, s_n$, and $P(w_i|l_i[\mathrm{h}(h, [s_1, \ldots, s_n])])$ is the probability of a lexical derivation $l_i \rightarrow w_i$, given that w_i functions as a slot-marker of lexical head $\mathrm{h}(h, [s_1, \ldots, s_n])$.

Finally, we estimate the probability of deriving each slot-filler, e.g. "*kanojo* (she)" and "*pai* (pie)", in assuming that the derivation of a slot-filler depends only on its head word and slot:

$$P(kanojo|R, ga, o, tabe, ta) \approx P(kanojo|N[\mathrm{s}(tabe, ga)]) \qquad (7)$$

$$P(pai|R, kanojo, ga, o, tabe, ta) \approx P(pai|N[\mathrm{s}(tabe, o)]) \qquad (8)$$

where $\mathrm{s}(h, s)$ is a lexical context denoting a slot s of a head word h, and $P(w_i|l_i[\mathrm{s}(h, s)])$ is the

does not impose any restriction on the representation of syntactic structures.

probability of a lexical derivation $l_i \rightarrow w_i$ given that w_i functions as a filler of a slot $\mathrm{s}(h, s)$.

Combining equations (3), (4), (5), (6), (7) and (8), we produce (9):

$$\begin{aligned} P(W|R) \;\approx\; & P(ta|Aux) \cdot P(tabe|V) \cdot \\ & P(ga|P[\mathrm{h}(tabe, [P, P])]) \cdot \\ & P(o|P[\mathrm{h}(tabe, [P\!:\!ga, P])]) \cdot \\ & P(kanojo|N[\mathrm{s}(tabe, ga)]) \cdot \\ & P(pai|N[\mathrm{s}(tabe, o)]) \qquad (9) \end{aligned}$$

2.3 Handling multiple lexical contexts

Note that a lexical derivation may be associated with more than one lexical context (multiple lexical contexts). Multiple lexical contexts appear typically in coordinate structures. For example, in the sentence shown in Figure 2, "*kanojo-wa* (she-TOP)" functions as the case of both of the verbs "*tabe* (eat)" and "*dekake* (leave)".

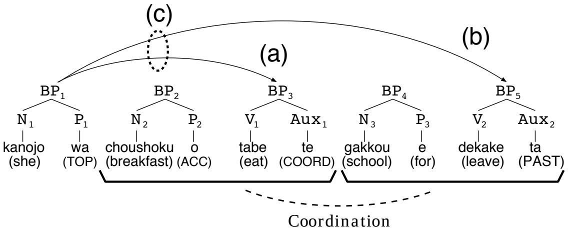

Figure 2: An example sentence containing a coordinate structure: "She ate breakfast and left for school"

Let us first consider the lexical derivation probability for the slot-filler "*kanojo* (she)". According to the assumption mentioned in Section 2.2, the lexical contexts of this slot-filler should be $\mathrm{s}(tabe, wa)$ and $\mathrm{s}(dekake, wa)$. Thus, the probability of deriving it is $P(kanojo|N_1[\mathrm{s}(tabe, wa), \mathrm{s}(dekake, wa)])$. More generally, if a slot-filler w_i is associated with two lexical contexts c_1 and c_2, then the probability of deriving w_i can be estimated as follows:

$$\begin{aligned} & P(w_i|l_i[c_1, c_2]) \\ & \quad = \frac{P(l_i[c_1, c_2]|w_i) \cdot P(w_i)}{P(l_i[c_1, c_2])} && (10) \\ & \quad \approx \frac{P(l_i[c_1]|w_i) \cdot P(l_i[c_2]|l_i, w_i) \cdot P(w_i)}{P(l_i[c_1]) \cdot P(l_i[c_2]|l_i])} && (11) \\ & \quad = P(w_i|l_i) \cdot \frac{P(w_i|l_i[c_1])}{P(w_i|l_i)} \cdot \frac{P(w_i|l_i[c_2])}{P(w_i|l_i)} && (12) \\ & \quad = P(w_i|l_i) \cdot D(w_i|l_i[c_1]) \cdot D(w_i|l_i[c_2]) && (13) \end{aligned}$$

In (13), we assume that the two lexical contexts c_1 and c_2 are mutually independent given l_i (and w_i):

$$P(l_i[c_2]|l_i[c_1]) \approx P(l_i[c_2]|l_i) \qquad (14)$$

$$P(l_i[c_2]|l_i[c_1], w_i) \approx P(l_i[c_2]|l_i, w_i) \qquad (15)$$

$D(w_i|l_i[c])$ is what we call a lexical dependency parameter, which is given by:

$$D(w_i|l_i[c]) = \frac{P(w_i|l_i[c])}{P(w_i|l_i)} \qquad (16)$$

$D(w_i|l_i[c])$ measures the degree of the dependency between the lexical derivation $l_i \rightarrow w_i$ and its lexical context c. It is close to one if w_i and c are highly independent. It becomes greater than one if w_i and c are positively correlated, whereas it becomes less than one and close to zero if w_i and c are negatively correlated. Thus, if we set a lexical dependency parameter to one, that means we create a model that neglects the dependency associated with that parameter. For example, the probability of deriving "*kanojo* (she)" in Figure 2 is calculated as follows.

$$\begin{aligned} & P(kanojo|N_1[\mathrm{s}(tabe, wa), \mathrm{s}(dekake, wa)]) \\ & \approx P(kanojo|N_1) \cdot D(kanojo|N_1[\mathrm{s}(tabe, wa)]) \\ & \quad \cdot D(kanojo|N_1[\mathrm{s}(dekake, wa)]) \qquad (17) \end{aligned}$$

Let us then move to the estimation of the probability of deriving the slot-markers "*wa* (TOP)", "*o* (ACC)", and "*e* (for)", where "*wa*" is associated with both "*tabe* (eat)" and "*dekake* (leave)", while "*o*" is associated only with "*tabe*", and "*ni*" is associated only with "*dekake*". To be mode general, let slot-marker w_0 is associated with two lexical contexts c_1 and c_2, and slot-markers w_1 and w_2 are, respectively, associated with c_1 and c_2. Assuming that w_1 and w_2 are mutually dependent, being both dependent on w_0, and c_1 and c_2 are mutually independent, the joint probability of the derivations of w_0, w_1 and w_2 can be estimated as (20) in Figure 3, similar to (13). For example, the probability of deriving "*wa* (TOP)", "*o* (ACC)", and "*e* (for)" in Figure 2 is calculated as (21) in Figure 3.

Summarizing equations (2), (13) and (16), the lexical model $P(W|R)$ can be estimated by the product of the context-free distribution of the lexical derivations $P_{cf}(W|L)$ and the degree of the dependency between the lexical derivations $D(W|R)$:

$$P(W|R) \approx P_{cf}(W|L) \cdot D(W|R) \qquad (22)$$

$$P_{cf}(W|L) = \prod_{i=1}^{m} P(w_i|l_i) \qquad (23)$$

$$D(W|R) = \prod_{i=1}^{m} \prod_{c \in C_{w_i}} D(w_i|l_i[c]) \qquad (24)$$

where C_{w_i} is the set of the lexical contexts of w_i.

$$\begin{aligned}
&P(w_0, w_1, w_2 | l_0[\mathrm{h}(h_1, [l_0, l_1]), \mathrm{h}(h_2, [l_0, l_2])], l_1[\mathrm{h}(h_1, [l_0, l_1])], l_2[\mathrm{h}(h_2, [l_0, l_2])]) \\
&\quad \approx \quad P(w_0 | l_0[\mathrm{h}(h_1, [l_0, l_1]), \mathrm{h}(h_2, [l_0, l_2])]) \cdot P(w_1 | l_1[\mathrm{h}(h_1, [l_0 \!:\! w_0, l_1])]) \cdot P(w_2 | l_2[\mathrm{h}(h_2, [l_0 \!:\! w_0, l_2])]) && (18) \\
&\quad \approx \quad P(w_0 | l_0) \cdot \frac{P(w_0 | l_0[\mathrm{h}(h_1, [l_0, l_1])])}{P(w_0 | l_0)} \cdot \frac{P(w_0 | l_0[\mathrm{h}(h_2, [l_0, l_2])])}{P(w_0 | l_0)} \cdot \\
&\qquad\quad P(w_1 | l_1[\mathrm{h}(h_1, [l_0 \!:\! w_0, l_1])]) \cdot P(w_2 | l_2[\mathrm{h}(h_2, [l_0 \!:\! w_0, l_2])]) && (19) \\
&\quad = \quad P(w_0 | l_0) \cdot D(w_0 | l_0[\mathrm{h}(h_1, [l_0, l_1])]) \cdot D(w_0 | l_0[\mathrm{h}(h_2, [l_0, l_2])]) \cdot \\
&\qquad\quad P(w_1 | l_1) \cdot D(w_1 | l_1[\mathrm{h}(h_1, [l_0 \!:\! w_0, l_1])]) \cdot P(w_2 | l_2) \cdot D(w_2 | l_2[\mathrm{h}(h_2, [l_0 \!:\! w_0, l_2])]) && (20)
\end{aligned}$$

$$\begin{aligned}
&P(wa, o, e | P_1[\mathrm{h}(tabe, [P_1, P_2]), \mathrm{h}(dekake, [P_1, P_3])], P_2[\mathrm{h}(tabe, [P_1, P_2])], P_3[\mathrm{h}(dekake, [P_1, P_3])]) \\
&\quad \approx \quad P(wa | P_1) \cdot D(wa | P_1[\mathrm{h}(tabe, [P_1, P_2])]) \cdot D(wa | P_1[\mathrm{h}(dekake, [P_1, P_3])]) \cdot \\
&\qquad\quad P(o | P_2) \cdot D(o | P_2[\mathrm{h}(tabe, [P_1 : wa, P_2])]) \cdot P(e | P_3) \cdot D(e | P_3[\mathrm{h}(dekake, [P_1 : wa, P_3])]) && (21)
\end{aligned}$$

Figure 3: The joint probability of the derivations of slot-markers

2.4 Summary of our model

From equations (1) and (22), the overall distribution $P(R, W)$ can be decomposed as follows:

$$P(R, W) \approx P(R) \cdot P_{cf}(W|L) \cdot D(W|R) \qquad (25)$$

where the first term $P(R)$ reflects part-of-speech bigram statistics and structural preference, the second term $P_{cf}(W|L)$ reflects the occurrence of each word, and the third term $D(W|R)$ reflects lexical association. Thus, equation (25) suggests that our model integrates these types of statistics, while maintaining modularity of lexical association.

Figure 4 shows the factors of the P(R,W) for the sentence in Figure 1. In this figure:

1. $P(R)$ reflects the syntactic preference.
2. $P_{cf}(W|L)$, which consists of $P(kanojo|N)$, $P(ga|P)$ etc., reflects the occurrence of each word.
3. $D(W|R)$, which consists of $D(o|N[\mathrm{h}(tabe, [\,])])$, $D(pai|N[\mathrm{s}(tabe, ACC)])$ etc., reflects the lexical association statistics.

In this way, our modeling maintains the modularity of different statistics types.

The modularity of the lexical model facilitates parameter estimation. Although the syntactic model ideally requires *fully* bracketed training corpora, training it is expected to be manageable since the model's parameter space tends to be only a small part of the overall parameter space. The lexical association statistics, on the other hand, may have a much larger parameter space, and thus may require much larger amounts of training data, as compared to the syntactic model. However, since our lexical model can be trained independently of syntactic preference, one can train it using *partially* parsed tagged corpora, which can be produced at a lower cost (i.e. automatically), as well as fully bracketed corpora. In fact, we used both a full-bracketed corpus and a partially parsed corpus in our experiment.

3 A preliminary experiment

Let us first briefly describe some fundamental features of Japanese syntax. A Japanese sentence can be analyzed as a sequence of so-called *bunsetu* phrases (BPs, hereafter) as illustrated in Figure 1. A BP is a chunk of words consisting of a content word (noun, verb, adjective, etc.) accompanied by some function word(s) (postposition, auxiliary, etc.). For example, the BP "*kanojo-ga*" (BP_1) in Figure 1 consists of the noun "*kanojo* (she)" followed by the postposition "*ga* (NOM)", which functions as a slot-marker. The BP "*tabe-ta*" (BP_3), on the other hand, consists of the verb "*tabe* (eat)" followed by the auxiliary "*ta* (PAST)".

Given a sequence of BPs, one can recognize dependency relations between them as illustrated in Figure 1. In Japanese, if BP_i precedes BP_j, and BP_i and BP_j are in a dependency relation, then BP_i is always the modifier of BP_j, and we say "BP_i modifies BP_j." For example, in Figure 1, both BP_1 and BP_2 modify BP_3.

For the preliminary evaluation of our model, we restricted our focus only on the model's performance for structural disambiguation excluding morphological disambiguation. Thus, the task of the parser was restricted to determination of the dependency structure of an input sentence, which is given together with the specification of word

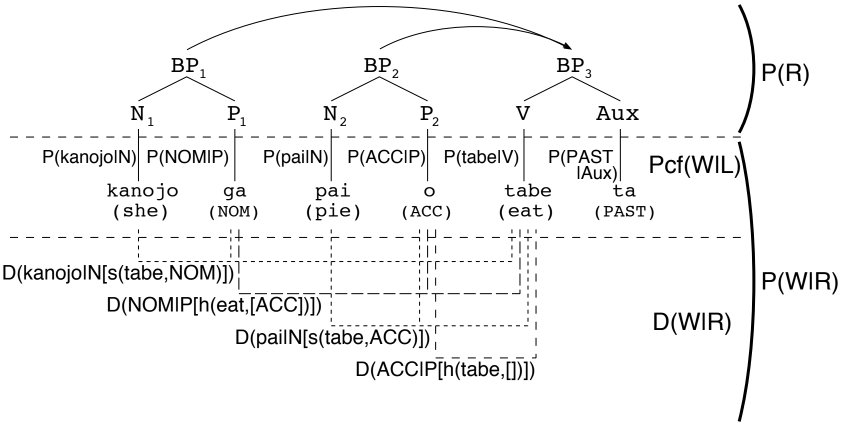

Figure 4: The summary of our model

segments, their POS tags, and the boundaries between BPs.

In developing the grammar used by our PGLR parser, we first established a categorization of BPs based on the POS of their constituents: postpositional BPs, verbal BPs, nominal predicative BPs, etc. We then developed a modification constraint matrix that describes which BP category can modify which BP category, based on examples collected from the Kyoto University text corpus [11]. We finally transformed this matrix into a CFG; for instance, the constraint that a BP of category C_i can modify a BP of category C_j can be transformed into context-free rules such as $\langle \bar{C}_j \rightarrow C_i \;\; C_j \rangle$, $\langle \bar{C}_j \rightarrow \bar{C}_i \;\; C_j \rangle$, etc., where $\bar{X}$ denotes a nonterminal symbol.

For the text data, we used roughly 10,000 sentences from the Kyoto University text corpus for training the syntactic model, and the whole EDR corpus [6] and the RWC POS-tagged corpus [16] for training the lexical model. For testing, we used 500 sentences collected from the Kyoto University text corpus with the average sentence length being 8.7 BPs. The data sets used for training and testing are mutually exclusive. The grammar used by our probabilistic GLR parser was a CFG automatically acquired from the training sentences, consisting of 967 context-free rules containing 50 nonterminal symbols and 43 terminal symbols (i.e. BP categories).

The baseline of the disambiguation performance was assessed by way of a naive strategy which selects the nearest possible modifiee (similarly to the right association principle in English) under the non-crossing constraint. The performance of this naive strategy was 62.4% in BP-based accuracy, where BP-based accuracy is the ratio of the number of the BPs whose modifiee is correctly identified to the total number of BPs (excluding the two rightmost BPs for each sentence). On the other hand, the syntactic model $P(R)$ achieved 72.1% in BP-based accuracy, 9.7 points above the baseline.

4 The contribution of the lexical model

In our experiment, we considered the following three lexical dependency parameters in the lexical model.

First, we considered the dependencies between slot-markers and their lexical head by using the lexical dependency parameter (26).

$$D(p|P[\mathrm{h}(h,[s_1,\ldots,s_n])]) \qquad (26)$$

(26) can be computed from $P(p^n|P^n[\mathrm{h}(h,[])])$, the distribution of n postpositions (slot-markers) given that all of them are subordinated by a single lexical head h. We trained this distribution using 150,000 instances of p^n-$\{verb,adjective,nominal\_predicate\}$ collocation collected from the EDR full-bracketed corpus. For parameter estimation, we used the maximum entropy estimation technique [1, 15]. For further details of this estimation process, see [20].

Next, we considered dependencies between slot-fillers and their head verb coupled with the corresponding slot-markers by using the lexical dependency parameter (27).

$$D(n|N[s(v,p)]) \qquad (27)$$

(27) was trained using 6.7 million instances of *noun-postposition-verb* collocation collected from both the EDR and RWC corpora. For parameter estimation, we used 115 non-hierarchical semantic noun classes derived from the NTT semantic

dictionary [7] to reduce the parameter space:

$$D(n|N[\mathrm{s}(v,p)]) \approx \frac{\sum_{c_n} P(c_n|N[\mathrm{s}(v,p)]) \cdot P(n|c_n)}{P(n|N)} \tag{28}$$

$P(c_n|N[\mathrm{s}(v,p)])$ was estimated using a simple back-off smoothing technique: for any given lexical verb v and postposition p, if the frequency of $\mathrm{s}(v,p)$ is less than a certain threshold λ (in our experiment, $\lambda = 100$), then $P(c_n|N[\mathrm{s}(v,p)])$ was approximated to be $P(c_n|N[\mathrm{s}(c_v,p)])$ where c_v is a class of v whose frequency is more than λ.

Finally, we considered the occurrence of postpositions by using the lexical dependency parameter (29).

$$D(p|P[head\_type]) \tag{29}$$

In Japanese, the distribution of the lexical derivation of postpositions, $P(p|P)$, is quite different depending on whether they function as slot-markers of verbs, adjectives and nominal predicates such as "*ga* (NOM)" and "*o* (ACC)" in Figure 1, or they function as slot-markers of nouns such as "*no* (of)" in the following sentence.

hana	*no*	*syashin*[3]
(flower)	(of)	(picture)

For such a reason, we introduced the lexical dependency parameter (29), where *head_type* denotes whether the postposition P functions as a slot-marker of a predicate or a noun. We estimated this dependency parameter using about 950,000 postpositions collected from the EDR corpus.

Table 1 summarizes the results of the experiment. The lexical model achieved 76.5% in BP-based accuracy, and the model using both the syntactic and lexical model achieved 82.8% in BP-based accuracy. According to these results, the contribution of lexical statistics for disambiguation is as great as that of syntactic statistics in our framework.

The bottom three lines in Table 1 denotes the setting where the only lexical dependency parameter (26), (27) and (29) are considered in the lexical model. Among these, the contribution of (29) was greatest.

5 Error analysis

In the test set, there were 574 BPs whose modifiee was not correctly identified by the system. Among these errors, we particularly explored 290 errors that were associated with postpositional BPs functioning as a case of either a verb, adjective, or nominal predicate, since, for lexical association statistics in the lexical model, we took the dependencies between slots (i.e. slot-markers and slot-fillers) and their heads into account. In this exploration, we identified three major error types: (a) errors associated with a coordinate clause, (b) errors associated with relative clauses, (c) errors associated with the lack of the consideration of dependency between slot-fillers.

Table 1: The contribution of the lexical model

	accuracy
base line	62.4 %
syntactic model only	72.1 %
lexical model only	76.5 %
syntactic + lexical model	82.8 %
syntactic model + (26)	73.4 %
syntactic model + (27)	78.3 %
syntactic model + (29)	81.3 %

5.1 Coordinate structures

One of the typical error types is associated with coordinate structures. The sentence in Figure 2 has at least three alternative interpretations in terms of which BP is modified by the left-most BP "*kanojo-wa* (she-TOP)": (a) "*tabe-ta* (eat-PAST)", (b) "*dekake-ta* (leave-PAST)", (c) both "*tabe-ta* (eat-PAST)" and "*dekake-ta* (leave-PAST)". Among these alternatives, the most reasonable interpretation is obviously (c), where the two predicative BPs constitute a coordinate structure.

In our experiment, however, neither the training data nor the test data indicates such coordinate structures. Thus, in the above sentence, for example, the system was required to choose one of two alternatives (a) and (b), where (b) is the preferred candidate according to the structural policy underlying our corpora. However, this choice is not really meaningful. Furthermore, the system systematically prefers (a), the wrong choice, since (i) the syntactic model tends to prefer shorter-distance modification relations (similarly to the right association principle in English), and (ii) the lexical model is expected to support both candidates because both $D(kanojo|N[\mathrm{s}(tabe, wa)])$ in (a) and $D(kanojo|N[\mathrm{s}(dekake, wa)])$ in (b) should be high. This problem makes the performance of our model lower than what it should be.

Obviously, the first step to resolving this problem is to enhance our corpora and grammar to enable the parser to generate the third interpretation, i.e. to explicitly generate a coordinate structure such as (c) if needed. Once such a setting is established, we then need to consider the

[3] This sentence means "a picture of a flower."

lexical contexts of each of the constituents modifying a coordinate structure, such as "*kanojo-wa* (she-TOP)" in the above sentence. In interpretation (c), since "*kanojo-wa* (she-TOP)" modifies both predicative BPs, it is reasonable to associate it with two lexical contexts, $\mathrm{s}(tabe, wa)$ and $\mathrm{s}(dekake, wa)$. As mentioned in Section 2, our framework allows us to deal with such multiple lexical contexts, namely:

$$\begin{aligned} &D(kanojo|N[\mathrm{s}(tabe, wa), \mathrm{s}(dekake, wa)]) \\ &\quad \approx D(kanojo|N[\mathrm{s}(tabe, wa)]) \cdot \\ &\qquad D(kanojo|N[\mathrm{s}(dekake, wa)]) \end{aligned} \tag{30}$$

The correct interpretation (c) would assigned higher probability than (a) or (b), since the two lexical dependency parameters in (30), $D(kanojo|N[\mathrm{s}(tabe, wa)])$ and $D(kanojo|N[\mathrm{s}(dekake, wa)])$ are both expected to be sufficiently large.

5.2 Treatment of correference

One may have already noticed that the issue discussed above can be generalized as an issue associated with the treatment of correference in dependency structures. Namely, if a prepositional BP is correferred to by more than one clause as a participant, a naive treatment of this correference relation could require the parser to make a meaningless choice: which clause subordinates that BP. This problem in the treatment of correference is considered to cause a significant proportion of errors associated with relative/adverbial clauses or compound predicates. Such errors are expected to be resolvable through an extension of the model, as discussed in Section 5.1.

Let us briefly look at another example in Figure 5, where the matrix clause and relative clause correfer to the leftmost BP "*kanojo-wa* (she-TOP)", i.e. interpretation (c). Without any refined treatment of this correference relation, the parser would be required to make a meaningless choice between (a) and (b).

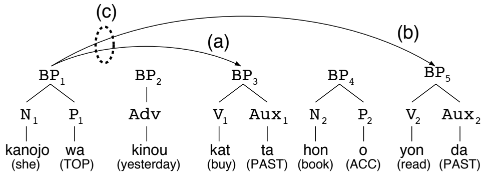

Figure 5: An example sentence containing a relative clause: "She read the book which she bought yesterday"

5.3 Dependency between slot fillers

According to the results summarized in Table 1, the contribution of the dependency between slot-fillers and their heads seems to be negligibly small. We can enumerate several possible reasons including that the estimation of these types of dependency parameters was not sufficiently sophisticated.

In addition to these reasons, we also found that the lack of the consideration of dependency between slot-fillers was also problematic in some cases; there are particular patterns where dependency between slot-fillers seems to be highly significant. For example, in the clause "*kanojo-wa* (she-TOP) *isha-ni* (doctor-DAT) *nat-ta* (become-PAST)" (she became a doctor), the distribution of the filler of the "*wa* (TOP)" slot is considered to be highly dependent on the filler of the "*ni* (DAT)" slot, "*isha* (doctor)", since its distribution would be markedly different if "*isha* (doctor)" was replaced with "*mizu* (water)". Similar patterns include, for example, "*A-wo* (ACC) *B-ni* (DAT) *suru* (make)", where A and B are highly dependent, and "*A-ga* (NOM) *B-wo* (ACC) *suru* (do)", where noun B indicating an action strongly influences the distribution of A.

In our framework, this type of problem can be treated by means of controlling the choice of lexical contexts. We are now conducting another experiment in which the dependencies between slot-fillers are additionally considered in particular patterns. Note that the refinement of our model in this manner illustrates that the modularity of lexical association statistics facilitates rule-based control in choosing the locations where lexical association is considered. This rule-based control allows us to incorporate qualitative knowledge such as linguistic insights and heuristics newly obtained from experiments based on the model.

6 Conclusion

In this paper, we first presented a new framework of language modeling for statistical parsing, which incorporates lexical association statistics while maintaining modularity. We then reported on the results of our preliminary evaluation of the model's performance, showing that both the syntactic and lexical models made a considerable contribution to structural disambiguation, and that the division of labor between those two models thus seemed to be working well to date.

Many issues remain unclear. First, we need to conduct experiments on the combination of the morphological and syntactic disambiguation tasks, which our framework intrinsically is designed for. Second, empirical comparison with other lexically sensitive models is also strongly required. One interesting issue is whether the division of labor between the syntactic and lexical models presented in this paper works well language-independently, or conversely, whether the existing models designed for English are equally applicable to languages like Japanese.

Acknowledgements

The authors would like to thank the staff of NTT for making available their considerable electronic resources.

References

[1] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, 1996.

[2] E. Black, F. Jelinek, J. Lafferty, D. M. Magerman, R. Mercer, and S. Roukos. Towards history-based grammars: Using richer models for probabilistic parsing. In *Proceedings of the ACL*, pages 31–37, 1993.

[3] E. Charniak. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the AAAI*, 1997.

[4] M. Collins. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the ACL*, 1996.

[5] M. Collins. Three generative, lexicalised models for statistical parsing. In *Proceedings of the ACL*, 1997.

[6] EDR. The EDR electronic dictionary technical guide (second edition). Technical Report TR–045, Japan Electronic Dictionary Research Institute, 1995.

[7] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi. *A Japanese Lexicon*. Iwanami Shoten, 1997. (In Japanese).

[8] K. Inui, K. Shirai, H. Tanaka, and T. Tokunaga. Integrated probabilistic language modeling for statistical parsing. Technical Report TR97-0005, Dept. of Computer Science, Tokyo Institute of Technology, 1997. `ftp://ftp.cs.titech.ac.jp/lab/tanaka/papers/97/inui97b.ps.gz`.

[9] K. Inui, K. Shirai, T. Tokunaga, and H. Tanaka. Integration of statistical techniques for parsing. In *summary collection of the IJCAI'97 poster session*, 1997.

[10] K. Inui, V. Sornlertlamvanich, H. Tanaka, and T. Tokunaga. A new formalization of probabilistic GLR parsing. In *Proceedings of the IWPT*, 1997.

[11] S. Kurohashi and M. Nagao. Kyoto university text corpus project. In *Proceedings of the 11th Annual Conference of JSAI*, pages 58–61, 1997. (In Japanese).

[12] H. Li. A probabilistic disambiguation method based on psycholinguistic principles. In *Proceedings of WVLC-4*, 1996.

[13] D. M. Magerman and M. Marcus. Pearl: A probabilistic chart parser. In *Proceedings of the EACL*, pages 15–20, 1991.

[14] D. M. Magerman. Statistical decision-tree models for parsing. In *Proceedings of the ACL*, pages 276–283, 1995.

[15] A. Ratnaparkhi, J. Reyner, and S. Roukos. A maximum entropy model for prepositional phrase attachment. In *Proceedings of the Human Language Technology Workshop*, pages 250–255, 1994.

[16] Real World Computing Partnership. RWC text database. `http://www.rwcp.or.jp/wswg.html`, 1995.

[17] P. Resnik. Probabilistic tree-adjoining grammar as a framework for statistical natural language processing. In *Proceedings of the COLING*, pages 418–424, 1992.

[18] Y. Schabes. Stochastic lexicalized tree-adjoining grammars. In *Proceedings of the COLING*, pages 425–432, 1992.

[19] S. Sekine and R. Grishman. A corpus-based probabilistic grammar with only two non-terminals. In *Proceedings of the IWPT*, 1995.

[20] K. Shirai, K. Inui, T. Tokunaga, and H. Tanaka. Learning dependencies between case frames using maximum entropy method. In *Proceedings of Annual Meeting of the Japan Association for Natural Language Processing*, 1997. (In Japanese).

[21] V. Sornlertlamvanich, K. Inui, K. Shirai, H. Tanaka, T. Tokunaga, and T. Takezawa. Empirical evaluation of probabilistic glr parsing. In *Proceedings of the NLPRS*, 1997.