Learning Probabilistic Subcategorization Preference by Identifying Case Dependencies and Optimal Noun Class Generalization Level*

Takehito Utsuro Yuji Matsumoto
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-01, JAPAN
{utsuro,matsu}@is.aist-nara.ac.jp

Abstract

This paper proposes a novel method of learning probabilistic subcategorization preference. In the method, for the purpose of coping with the ambiguities of case dependencies and noun class generalization of argument/adjunct nouns, we introduce a data structure which represents a tuple of independent partial subcategorization frames. Each collocation of a verb and argument/adjunct nouns is assumed to be generated from one of the possible tuples of independent partial subcategorization frames. Parameters of subcategorization preference are then estimated so as to maximize the subcategorization preference function for each collocation of a verb and argument/adjunct nouns in the training corpus. We also describe the results of the experiments on learning probabilistic subcategorization preference from the EDR Japanese bracketed corpus, as well as those on evaluating the performance of subcategorization preference.

1 Introduction

In corpus-based NLP, extraction of linguistic knowledge such as lexical/semantic collocation is one of the most important issues and has been intensively studied in recent years. In those research, extracted lexical/semantic collocation is especially useful in terms of ranking parses in syntactic analysis as well as automatic construction of lexicon for NLP.

For example, in the context of syntactic disambiguation, Black (1993) and Magerman (1995) proposed statistical parsing models based-on decision-tree learning techniques, which incorporated not only syntactic but also lexical/semantic information in the decision-trees. As lexical/semantic information, Black (1993) used about 50 semantic categories, while Magerman (1995) used lexical forms of words. Collins (1996) proposed a statistical parser which is based on probabilities of dependencies between head-words in the parse tree. In those works, lexical/semantic collocation are used for ranking parses in syntactic analysis.

On the other hand, in the context of automatic lexicon construction, the emphasis is mainly on the extraction of lexical/semantic collocational knowledge of specific words rather than its use in sentence parsing. For example, Haruno (1995) applied an information-theoretic data compression technique to corpus-based case frame learning, and proposed a method of finding case frames of verbs as compressed representation of verb-noun collocational data in corpus. The work concentrated on the extraction of declarative representation of case frames and did not consider their performance in sentence parsing.

This paper focuses on extracting lexical/semantic collocational knowledge of verbs for the purpose of applying it to ranking parses in syntactic analysis. More specifically, we propose a novel method for learning parameters for calculating subcategorization preference functions of verbs. In general, when learning lexical/semantic collocational knowledge of verbs from corpus, it is necessary to cope with the following two types of ambiguities:

1) *The ambiguity of case dependencies*
2) *The ambiguity of noun class generalization*

1) is caused by the fact that, only by observing each verb-noun collocation in corpus, it is not decidable which cases are dependent on each other and which cases are optional and independent of other cases. 2) is caused by the fact that, only by observing each verb-noun collocation in corpus, it is not decidable which superordinate class generates each observed leaf class in the verb-noun collocation.

So far, there exist several researches which worked on these two issues in learning collocational knowledge of verbs and also evaluated the results in terms of syntactic disambiguation. Resnik (1993) and Li and Abe (1995) studied how to find an optimal abstraction level of an argument noun in a tree-structured thesaurus. Although they evaluated the obtained abstraction level of the argument noun by its performance in syntactic disambiguation, their works are limited to only one argument. Li and Abe (1996) also studied a method for learning dependencies between case slots and evaluated the discovered dependencies in the syntactic disambiguation task. They first obtained optimal abstraction levels of the argument nouns by the method in Li and Abe (1995), and then tried to discover dependencies between the class-based case slots. They reported that dependencies

[*] The authors would like to thank Dr. Hang Li of NEC C&C Research Laboratories, Dr. Kentaro Inui of Tokyo Institute of Technology, Dr. Koiti Hasida of Electrotechnical Laboratory, Dr. Takashi Miyata of Nara Institute of Science and Technology, and also anonymous reviewers of ANLP97 for valuable comments on this work.

were discovered only at the slot-level and not at the class-level.

Compared with those previous works, this paper proposes to cope with the above two ambiguities in a uniform way. First, we introduce a data structure which represents a tuple of independent partial subcategorization frames. Each collocation of a verb and argument/adjunct nouns is assumed to be generated from one of the possible tuples of independent partial subcategorization frames. Then, parameters of subcategorization preference are estimated so as to maximize the subcategorization preference function for each collocation of a verb and argument/adjunct nouns in the training corpus. We describe the results of the experiments on learning probabilistic subcategorization preference from the EDR Japanese bracketed corpus (EDR, 1995), as well as those on evaluating the performance of subcategorization preference.

2 Data Structure

2.1 Verb-Noun Collocation

Verb-noun collocation is a data structure for the collocation of a verb and all of its argument/adjunct nouns. A verb-noun collocation e is represented by a feature structure which consists of the verb v and all the pairs of co-occurring case-markers p and thesaurus classes c of case-marked nouns:[1]

$$e \;=\; \begin{bmatrix} pred : v \\ p_1 : c_1 \\ \vdots \\ p_k : c_k \end{bmatrix} \qquad (1)$$

We assume that a *thesaurus* is a tree-structured type hierarchy in which each node represents a semantic class, and each thesaurus class $c_1, \ldots, c_k$ in a verb-noun collocation is a leaf class. We also introduce $\preceq_c$ as the superordinate-subordinate relation of classes in a thesaurus: $c_1 \preceq_c c_2$ means that c_1 is subordinate to c_2.

2.2 Subcategorization Frame

A *subcategorization frame* f is represented by a feature structure which consists of a verb v and the pairs of case-markers p and sense restriction c of case-marked argument/adjunct nouns:

$$f \;=\; \begin{bmatrix} pred : v \\ p_1 : c_1 \\ \vdots \\ p_l : c_l \end{bmatrix}$$

Sense restriction $c_1, \ldots, c_l$ of case-marked argument/adjunct nouns are represented by classes at arbitrary levels of the thesaurus. A subcategorization frame f can be divided into two parts: one is the verbal part f_v containing the verb v while the other is the nominal part f_p containing all the pairs of case-markers p and sense restriction c of case-marked nouns.

$$f \;=\; f_v \wedge f_p \;=\; \begin{bmatrix} pred : v \end{bmatrix} \wedge \begin{bmatrix} p_1 : c_1 \\ \vdots \\ p_l : c_l \end{bmatrix}$$

2.3 Subsumption Relation

We introduce *subsumption relation* $\preceq_f$ of a *verb-noun collocation* e and a *subcategorization frame* f:

$e \preceq_f f$ iff. for each case-marker p_i in f and its noun class c_{if}, there exists the same case-marker p_i in e and its noun class c_{ie} is subordinate to c_{if}, i.e. $c_{ie} \preceq_c c_{if}$

The subsumption relation $\preceq_f$ is applicable also as a subsumption relation of two subcategorization frames.

3 A Model of Generating Verb-Noun Collocation

In this section, we introduce a model of generating a verb-noun collocation from subcategorization frame(s). In order to cope with the ambiguities of *case dependencies* and *noun class generalization* in this model, we introduce a data structure which represents a tuple of *independent partial subcategorization frames*.

3.1 Generating a Verb-Noun Collocation from Independent Partial Subcategorization Frames

First, we describe the idea of generating a verb-noun collocation from a subcategorization frame, or a tuple of partial subcategorization frames.

Generation from a Subcategorization Frame

Suppose a verb-noun collocation e is given as:

$$e \;=\; \begin{bmatrix} pred : v \\ p_1 : c_{1e} \\ \vdots \\ p_k : c_{ke} \end{bmatrix}$$

Then, let us consider a subcategorization frame f which can generate e. We assume that f has exactly the same case-markers as e has,[2] and each semantic class c_{if} of a case-marked noun of f is superordinate to the corresponding leaf semantic class c_{ie} of e:

$$f \;=\; \begin{bmatrix} pred : v \\ p_1 : c_{1f} \\ \vdots \\ p_k : c_{kf} \end{bmatrix}, \quad c_{ie} \preceq_c c_{if} \; (i=1,\ldots,k) \qquad (2)$$

Then, we denote the generation of the verb-noun collocation e from the subcategorization frame f as:

$$f \;\longrightarrow\; e$$

Next, we describe the idea of generating a verb-noun collocation from a tuple of partial subcategorization frames which are independent of each other.

[1] Although we ignore sense ambiguities of case-marked nouns in this definition, in section 5.2, we briefly mention how we deal with sense ambiguities of case-marked nouns in the current implementation.

[2] Since we do not consider ellipsis of argument nouns when generating a verb-noun collocation from a subcategorization frame, the subcategorization frame f is required to have exactly the same case-markers as e.

Partial Subcategorization Frame

First, we define a *partial subcategorization frame* f_i of f as a subcategorization frame which has the same verb v as f as well as some of the case-markers of f and their semantic classes. Then, we can find a division of f into a tuple $\langle f_1, \ldots, f_n \rangle$ of partial subcategorization frames of f, where any pair f_i and $f_{i'}$ $(i \neq i')$ do not have common case-markers and the unification $f_1 \wedge \cdots \wedge f_n$ of all the partial subcategorization frames equals to f:

$$f \;=\; f_1 \wedge \cdots \wedge f_n \qquad (3)$$

$$f_i \;=\; \begin{bmatrix} pred : v \\ \vdots \\ p_{ij} : c_{ij} \\ \vdots \end{bmatrix}, \quad \begin{array}{l} \forall j \forall j' \; p_{ij} \neq p_{i'j'} \\ (i, i' = 1, \ldots, n, \quad i \neq i') \end{array} \qquad (4)$$

Independence of Partial Subcategorization Frames

We allow the division of f into a tuple $\langle f_1, \ldots, f_n \rangle$ of partial subcategorization frames as in the equation (3) only when the partial subcategorization frames $f_1, \ldots, f_n$ can be regarded as events occurring *independently* of each other. With some corpus, usually we can estimate the conditional probabilities $p(f \mid v)$ and $p(f_i \mid v)$ of the (partial) subcategorization frames f and f_i $(i = 1, \ldots, n)$ given the verb v. According to the estimated probabilities, we can judge whether $f_1, \ldots, f_n$ are *independent* of each other as follows.

First, we estimate the conditional probability $p(f \mid v)$ of a (partial) subcategorization frame f by summing up the conditional probabilities $p(e \mid v)$ of all the verb-noun collocations e given the verb v, where e is subsumed by f $(e \preceq_f f)$[3]

$$p(f \mid v) \;\approx\; \sum_{e \preceq_f f} p(e \mid v) \qquad (5)$$

The conditional joint probability $p(f_1, \ldots, f_n \mid v)$ is also estimated by summing up $p(e \mid v)$ where e is subsumed by all of $f_1, \ldots, f_n$ $(e \preceq_f f_1, \ldots, f_n)$:

$$p(f_1, \ldots, f_n \mid v) \;\approx\; \sum_{e \preceq_f f_1, \ldots, f_n} p(e \mid v) \qquad (6)$$

Then, we give a formal definition of *independence* of partial subcategorization frames according to the estimated conditional probabilities:

> partial subcategorization frames $f_1, \ldots, f_n$ are *independent* if, any pair f_i and f_j $(i \neq j)$ do not have common case-markers, and for every subset $f_{i_1}, \ldots, f_{i_j}$ of j of these partial subcategorization frames $(j = 2, \ldots, n)$, the following equation holds:

$$p(f_{i_1}, \ldots, f_{i_j} \mid v) \;=\; p(f_{i_1} \mid v) \cdots p(f_{i_j} \mid v) \qquad (7)$$

Since it is too strict to judge the independence of partial subcategorization frames by the equation (7), we relax the constraint of independence using a relaxation parameter α $(0 \leq \alpha \leq 1)$. Partial subcategorization frames $f_1, \ldots, f_n$ are judged as *independent* if, for every subset $f_{i_1}, \ldots, f_{i_j}$ of j of these partial subcategorization frames $(j = 2, \ldots, n)$, the following inequalities hold:

$$\alpha \;\leq\; \frac{p(f_{i_1}, \ldots, f_{i_j} \mid v)}{p(f_{i_1} \mid v) \cdots p(f_{i_j} \mid v)} \;\leq\; \frac{1}{\alpha} \qquad (8)$$

Generation from Independent Partial Subcategorization Frames

Now, as in the case of the generation from a subcategorization frame f, we denote the generation of e from a tuple $\langle f_1, \ldots, f_n \rangle$ of independent partial subcategorization frames of f as below:

$$\langle f_1, \ldots, f_n \rangle \;\longrightarrow\; e$$

3.2 The Ambiguity of Case Dependencies

This section describes the problem of the ambiguity of case dependencies when observing verb-noun collocation in corpus. This problem is caused by the fact that, only by observing each verb-noun collocation in corpus, it is not decidable which cases are dependent on each other and which cases are optional and independent of other cases.

For example, consider the following example:

Example 1

Kodomo-ga	kouen-de	juusu-wo	nomu.
child-NOM	*park-at*	*juice-ACC*	*drink*

(A child drinks juice at the park.)

The verb-noun collocation is represented as a feature structure e below:

$$e \;=\; \begin{bmatrix} pred : nomu \\ ga : c_c \\ wo : c_j \\ de : c_p \end{bmatrix}$$

In this feature structure e, c_c, c_p, and c_j represent the leaf classes (in the thesaurus) of the nouns *"kodomo(child)"*, *"kouen(park)"*, and *"juusu(juice)"*.

Next, we assume that the concepts *"human"*, *"place"*, and *"beverage"* are superordinate to *"kodomo(child)"*, *"kouen(park)"*, and *"juusu(juice)"*, respectively, and introduce the corresponding classes c_{hum}, c_{plc}, and c_{bev}. Then, the following superordinate-subordinate relations hold:

$$c_c \preceq_c c_{hum}, \quad c_p \preceq_c c_{plc}, \quad c_j \preceq_c c_{bev}$$

Allowing these superordinate classes as sense restriction in subcategorization frames, let us consider the several patterns of subcategorization frames which can generate the verb-noun collocation e. Those patterns of subcategorization frames vary according to the dependencies of cases within them.

If the three cases *"ga(NOM)"*, *"wo(ACC)"*, and *"de(at)"* are dependent on each other and it is not possible to find any division into a tuple of several independent partial subcategorization frames, e can be regarded as generated from a subcategorization frame containing all of the three cases:

$$\left\langle \begin{bmatrix} pred : nomu \\ ga : c_{hum} \\ wo : c_{bev} \\ de : c_{plc} \end{bmatrix} \right\rangle \;\longrightarrow\; e \qquad (9)$$

[3] The probability $p(e \mid v)$ can be estimated as $freq(e)/freq(v)$ by M.L.E. (maximum likelihood estimation) directly from the training corpus.

Otherwise, if only the two cases *"ga(NOM)"* and *"wo(ACC)"* are dependent on each other and the *"de(at)"* case is independent of those two cases, e can be regarded as generated from the following tuple of independent partial subcategorization frames:

$$\left\langle \begin{bmatrix} pred : nomu \\ ga : c_{hum} \\ wo : c_{bev} \end{bmatrix}, \begin{bmatrix} pred : nomu \\ de : c_{plc} \end{bmatrix} \right\rangle \longrightarrow e \tag{10}$$

Otherwise, if all the three cases *"ga(NOM)"*, *"wo(ACC)"*, and *"de(at)"* are independent of each other, e can be regarded as generated from the following tuple of independent partial subcategorization frames, each of which contains only one case:

$$\left\langle \begin{bmatrix} pred : nomu \\ ga : c_{hum} \end{bmatrix}, \begin{bmatrix} pred : nomu \\ wo : c_{bev} \end{bmatrix}, \begin{bmatrix} pred : nomu \\ de : c_{plc} \end{bmatrix} \right\rangle \longrightarrow e \tag{11}$$

3.3 The Ambiguity of Noun Class Generalization

This section describes the problem of the ambiguity of noun class generalization when observing verb-noun collocation in corpus. This problem is caused by the fact that, only by observing each verb-noun collocation in corpus, it is not decidable which superordinate class generates each observed leaf class in the verb-noun collocation.

For example, let us again consider Example 1. We assume that the concepts *"animal"* and *"liquid"* are superordinate to *"human"* and *"beverage"*, respectively, and introduce the corresponding classes c_{ani} and c_{liq}. Then, the following superordinate-subordinate relations hold:

$$c_{hum} \preceq_c c_{ani}, \quad c_{bev} \preceq_c c_{liq}$$

If we additionally allow these superordinate classes as sense restriction in subcategorization frames, we can consider several additional patterns of subcategorization frames which can generate the verb-noun collocation e, along with those patterns described in the previous section.

Suppose that only the two cases *"ga(NOM)"* and *"wo(ACC)"* are dependent on each other and the *"de(at)"* case is independent of those two cases as in the formula (10). Since the leaf class c_c (*"child"*) can be generated from either c_{hum} or c_{ani}, and also the leaf class c_j (*"juice"*) can be generated from either c_{bev} or c_{liq}, e can be regarded as generated according to either of the four formulas (10) and (12)~(14):

$$\left\langle \begin{bmatrix} pred : nomu \\ ga : c_{ani} \\ wo : c_{bev} \end{bmatrix}, \begin{bmatrix} pred : nomu \\ de : c_{plc} \end{bmatrix} \right\rangle \longrightarrow e \tag{12}$$

$$\left\langle \begin{bmatrix} pred : nomu \\ ga : c_{hum} \\ wo : c_{liq} \end{bmatrix}, \begin{bmatrix} pred : nomu \\ de : c_{plc} \end{bmatrix} \right\rangle \longrightarrow e \tag{13}$$

$$\left\langle \begin{bmatrix} pred : nomu \\ ga : c_{ani} \\ wo : c_{liq} \end{bmatrix}, \begin{bmatrix} pred : nomu \\ de : c_{plc} \end{bmatrix} \right\rangle \longrightarrow e \tag{14}$$

3.4 A Model of Generating Verb-Noun Collocation

When observing each verb-noun collocation e, as we described in the previous two sections, the ambiguities of case dependencies and noun class generalization remain, and it is necessary to consider every possible tuple of independent partial subcategorization frames which can generate the observed verb-noun collocation e. In order to cope with these ambiguities, we introduce two sets: one is a set $\mathbf{F}$ of tuples $\langle f_1, \ldots, f_n \rangle$ of independent partial subcategorization frames and the other is a set $\mathbf{E}$ of verb-noun collocations e. The generation of a verb-noun collocation from a tuple of independent partial subcategorization frames can be regarded as a mapping π from $\mathbf{F}$ to $\mathbf{E}$:

$$\pi : \mathbf{F} \rightarrow \mathbf{E} \tag{15}$$

Usually, for each given verb-noun collocation in $\mathbf{E}$, there exist several possible tuples of independent partial subcategorization frames in $\mathbf{F}$. Thus, π is a many-to-one mapping. The mapping from a tuple $\langle f_1, \ldots, f_n \rangle$ of independent partial subcategorization frames to a verb-noun collocation e can be denoted also as follows:

$$\langle f_1, \ldots, f_n \rangle \longrightarrow e \tag{16}$$

When observing a verb-noun collocation e, we assume this many-to-one mapping π and consider every possible tuple of independent partial subcategorization frames which can generate e, according to the ambiguities of case dependencies and noun class generalization.

3.5 Parameters of Generating Verb-Noun Collocation

Before we give definitions of subcategorization preference functions in the next section, we introduce *the parameter* $q(f_k \mid v)$ of generating verb-noun collocation, which is used in the calculation of the subcategorization preference. The parameter $q(f_k \mid v)$ can be regarded as the conditional probability of the partial subcategorization frame f_k and could be estimated in the similar way as the $p(f \mid v)$ in the formula (5). However, it is the parameter of *generating verb-noun collocation* and have to be estimated so as to maximize the subcategorization preference function for the training corpus.

One solution of this parameter estimation process might be to regard the model of generating verb-noun collocation as a probabilistic model and then to apply the maximum likelihood estimation method. When estimating the parameters from the training sample, we have to note that each verb-noun collocation is ambiguous since it could be interpreted in several different ways according to case dependencies and optimal noun class generalization levels. As for parameter estimation of probabilistic models from ambiguous training sample, EM algorithm(Baum, 1972) is a well-known solution and has been studied for years. In EM algorithm, parameters are assigned to events, and it is required that parameters sum up to 1. However, since two subcategorization frames could have the same case and a subsumption relation could hold between their sense restrictions, they may have overlap and the requirement that parameters sum up to 1 is not satisfiable. Therefore, it is not so straightforward to apply EM algorithm to the task of parameter estimation of generating verb-noun collocation.

Instead of introducing a probabilistic model of generating verb-noun collocation[4], in this paper, we employ more general framework which is applicable to various measures of subcategorization preference including the probability of generating verb-noun collocation. In the framework, the process of parameter estimation is regarded as a general optimization problem of maximizing the subcategorization preference function for the training corpus.

In order to describe the framework, first we introduce the probability $p(\langle f_1,\ldots,f_n\rangle_j \rightarrow e_i \mid e_i)$ of generating a verb-noun collocation e_i in the set $\mathbf{E}$ from a tuple $\langle f_1,\ldots,f_n\rangle_j$ in the set $\mathbf{F}$, given e_i, and denote it as a conditional probability $p(\langle f_1,\ldots,f_n\rangle_j \mid e_i)$. Then, for each e_i in $\mathbf{E}$, we can consider a probability distribution $p(\langle f_1,\ldots,f_n\rangle_j \mid e_i)$ over the set $\mathbf{F}$ of tuples of independent partial subcategorization frames:

			$\mathbf{E}$	
		e_1	$\cdots$	e_l
$\mathbf{F}$	$\langle f_1,\ldots,f_{n'}\rangle_1$		$\cdots$	
	$\vdots$	$\vdots$	$p(\langle f_1,\ldots,f_n\rangle_j \mid e_i)$	$\vdots$
	$\langle f_1,\ldots,f_{n''}\rangle_m$		$\cdots$	

Each probability distribution $p(\langle f_1,\ldots,f_n\rangle_j \mid e_i)$ satisfies the following axiom of the probability:

$$\sum_j p(\langle f_1,\ldots,f_n\rangle_j \mid e_i) \;=\; 1 \quad \text{for all } i$$

According to the probability distribution $p(\langle f_1,\ldots, f_n\rangle_j \mid e_i)$ of generating e_i from $\langle f_1,\ldots,f_n\rangle_j$, we estimate the frequency of the subcategorization frame f_k and then estimate the parameter $q(f_k \mid v)$ as below:

$$q(f_k \mid v) \approx \frac{freq(f_k)}{freq(v)} \approx \frac{\sum\limits_{i,j} 1\cdot p(\langle f_1,\ldots,f_k,\ldots,f_n\rangle_j \mid e_i)}{freq(v)} \tag{17}$$

When learning probabilistic subcategorization preference (section 5), we estimate the probability distribution $p(\langle f_1,\ldots,f_n\rangle_j \mid e_i)$ for each e_i so as to maximize the subcategorization preference function for e_i.

4 Subcategorization Preference Functions

This section introduces a function ϕ which measures the subcategorization preference when generating a verb-noun collocation e from a tuple $\langle f_1,\ldots,f_n\rangle$ of independent partial subcategorization frames:

$$\phi(\langle f_1,\ldots,f_n\rangle \longrightarrow e) \tag{18}$$

In this paper, we introduce a subcategorization preference function which is based-on the idea of *Kullback Leibler distance.*[5]

4.1 Nominal Parts of (Partial) Subcategorization Frames

First, let $f_p, f_{p_1},\ldots,f_{p_n}$ be the nominal parts of (partial) subcategorization frames $f, f_1,\ldots,f_n$ in the equations (2) and (4), respectively:

$$f_p \;=\; \begin{bmatrix} p_1 : c_{1f} \\ \vdots \\ p_k : c_{kf} \end{bmatrix}$$

$$f_p \;=\; f_{p_1} \wedge \cdots \wedge f_{p_n}$$

$$f_{p_i} \;=\; \begin{bmatrix} \vdots \\ p_{ij} : c_{ij} \\ \vdots \end{bmatrix}, \quad \begin{array}{l} \forall j \forall j' \; p_{ij} \neq p_{i'j'} \\ (i,i'=1,\ldots,n, \;\; i\neq i') \end{array}$$

As in the case of the parameters $q(f_i \mid v)$ of f_i given the verb v, we estimate the probability $p(f_{p_i})$ of the nominal part f_{p_i} in the whole corpus and call it *the parameter* $q(f_{p_i})$ *of* f_{p_i} *in the whole training corpus.* We estimate the frequency of f_{p_i} throughout the whole training corpus and then estimate the parameter $q(f_{p_i})$ of f_{p_i} as below:

$$\begin{aligned} q(f_{p_k}) &\approx \frac{\sum\limits_{v} freq(f_k)}{N} \\ &\approx \frac{\sum\limits_{v}\sum\limits_{i,j} 1\cdot p(\langle f_1,\ldots,f_k,\ldots,f_n\rangle_j \mid e_i)}{N} \end{aligned} \tag{19}$$

4.2 ϕ_{kl}: Kullback Leibler Distance

Rather than the simple conditional probability, this preference function is intended to measure the information-theoretic association of the verb v and the nominal part of the subcategorization frame.

The Kullback Leibler (KL) distance is a measure of the distance between two probability distribution. Given a random variable $\mathbf{X}$ and two probability distributions $p(\mathbf{X})$ and $q(\mathbf{X})$, the KL distance $\mathrm{D}(p\|q)$ of $p(\mathbf{X})$ and $q(\mathbf{X})$ is defined as below(Cover and Thomas, 1991), where each term can be regarded as the distance of two probabilities $p(x)$ and $q(x)$ of an event x:

$$\mathrm{D}(p\|q) \;=\; \sum_{x\in\mathbf{X}} p(x)\log\frac{p(x)}{q(x)}$$

In order to apply the idea of the KL distance to measuring the association of the verb v and the nominal part f_p of f, we introduce a random variable $\mathbf{F_p}$ which takes f_p as its value. We also introduce the probability distribution $p(\mathbf{F_p})$ of $\mathbf{F_p}$ and the conditional probability distribution $p(\mathbf{F_p} \mid v)$ of $\mathbf{F_p}$ given the verb v. Then, the KL distance of $p(\mathbf{F_p} \mid v)$ and $p(\mathbf{F_p})$ is denoted as $\mathrm{D}(p(\mathbf{F_p} \mid v)\|p(\mathbf{F_p}))$ and each term of it can be regarded as the distance of two probabilities $p(f_p \mid v)$ and $p(f_p)$. We assume that the larger this distance is, the stronger the association of f_p and v is, and measure the association of f_p and v with this

[4] Another alternative of solving the problem of learning probabilistic subcategorization preference based-on a probabilistic model is to regard the problem as the construction of probabilistic models from the training sample. We will discuss this issue in section 7.

[5] In Utsuro and Matsumoto (1997), we defined another subcategorization preference function ϕ_p which is based-on the *probability* of generating the verb-noun collocation and described experimental results of applying ϕ_p to the task of learning probabilistic subcategorization.

distance of the two probabilities $p(f_p \mid v)$ and $p(f_p)$. With this idea, the subcategorization preference function ϕ_{kl} is now formally defined as below:[6] [7]

$$
\begin{aligned}
&\phi_{kl}(\langle f_1,\ldots,f_n\rangle \longrightarrow e)\\
&= \; p(f_p \mid v)\log\frac{p(f_p \mid v)}{p(f_p)} \qquad (20)\\
&\approx \; \prod_{i=1}^{n} p(f_{p_i} \mid v) \times \log\frac{\prod_{i=1}^{n} p(f_{p_i} \mid v)}{\prod_{i=1}^{n} p(f_{p_i})} \qquad (21)\\
&\approx \; \prod_{i=1}^{n} q(f_{p_i} \mid v) \times \log\frac{\prod_{i=1}^{n} q(f_{p_i} \mid v)}{\prod_{i=1}^{n} q(f_{p_i})} \qquad (22)
\end{aligned}
$$

(21) is derived from the independence of the partial subcategorization frames $f_1,\ldots,f_n$. In (22), we use the parameters $q(f_{p_i} \mid v)$ and $q(f_{p_i})$ as an approximation of the probabilities $p(f_{p_i} \mid v)$ and $p(f_{p_i})$.

5 Learning Probabilistic Subcategorization Preference

The problem of learning subcategorization preference can be formalized as an optimization problem of estimating the probability distribution $p(\langle f_1,\ldots,f_n\rangle_j \mid e_i)$ (in section 3.5) of generating e_i from $\langle f_1,\ldots,f_n\rangle_j$ (and then the parameters $q(f_{p_k} \mid v)$ and $q(f_{p_k})$) so as to maximize the value of the subcategorization preference function for the whole training corpus. In this paper, we give only an approximate solution to this problem: we estimate the probability distribution $p(\langle f_1,\ldots,f_n\rangle_j \mid e_i)$ for each e_i so as to maximize the value of the subcategorization preference function *only* for e_i, not for the whole training corpus.

5.1 Problem Setting

Let the training corpus $\mathcal{E}$ be the set of verb-noun collocation e. We define the subcategorization preference $\hat{\phi}(e)$ of a verb-noun collocation e as the maximum of the subcategorization preference function ϕ (the formula (18)) of generating e from a tuple $\langle f_1,\ldots,f_n\rangle$.

$$\hat{\phi}(e) \;=\; \max_{\langle f_1,\ldots,f_n\rangle} \phi(\langle f_1,\ldots,f_n\rangle \longrightarrow e) \qquad (23)$$

Now, the problem of *learning probabilistic subcategorization preference* is stated as:

> for every verb-noun collocation e in $\mathcal{E}$, estimating the probability distribution $p(\langle f_1,\ldots,f_n\rangle_j \mid e)$ of generating e from $\langle f_1,\ldots,f_n\rangle_j$, under the constraint that the value of the subcategorization preference $\hat{\phi}(e)$ is *maximized.*

5.2 Learning Algorithm

First, we identify *independent* partial subcategorization frames according to the condition of (8). Then, let $E(v)$ be the set of verb-noun collocations containing the verb v in the training corpus $\mathcal{E}$. Let $F(e)$ be the set of tuples $\langle f_1,\ldots,f_n\rangle$ of independent partial subcategorization frames which can generate e and satisfy the independence condition of (8).[8]

$$F(e) \;=\; \Big\{\langle f_1,\ldots,f_n\rangle \Big| \langle f_1,\ldots,f_n\rangle \longrightarrow e\Big\} \qquad (24)$$

$F(e)$ contains a tuple $\langle f\rangle$ consisting of only one subcategorization frame f only if f can not be divided into several independent partial subcategorization frames.

Then, we assume that each element of $F(e)$ occurs evenly and estimate the initial conditional probability distribution $p(\langle f_1,\ldots,f_n\rangle_j \mid e)$ of generating e from $\langle f_1,\ldots,f_n\rangle_j$ as an approximation below:

$$p(\langle f_1,\ldots,f_n\rangle_j \mid e) \;\approx\; \frac{1}{|F(e)|} \qquad (25)$$

5.2.1 Approximate Estimation of Verb-Independent Parameters

Using the initial conditional probability distribution of $p(\langle f_1,\ldots,f_n\rangle_j \mid e)$ as in the formula (25), the initial values of the verb-independent parameters $q(f_{p_k})$ are estimated by the formulas (19). In the current implementation of the learning algorithm, we use these initial values as approximate estimation of those verb-independent parameters and probabilities throughout the learning process.

5.2.2 Iterative Reestimation of Verb-Dependent Parameters

Verb-dependent parameters $q(f_k \mid v)(= q(f_{p_k} \mid v))$ are iteratively estimated so as to maximize the subcategorization preference $\hat{\phi}(e)$ for every verb-noun collocation e in the training corpus $\mathcal{E}$. As a learning algorithm, we employ the following *stingy algorithm:*

1. Initialization

As with the case of the verb-independent parameters, for each verb-noun collocatoin e in $\mathcal{E}$, the set $F(e)$ is initially constructed according to the definition in (24). Then, the initial conditional probability distribution of $p(\langle f_1,\ldots,f_n\rangle_j \mid e)$ and the initial values of the verb-dependent parameters $q(f_k \mid v)$ are estimated as (25) and (17), respectively.

[6] Resnik (1993) applys the idea of the KL distance to measuring the association of a verb v and its object noun class c. Our definition of ϕ_{kl} corresponds to an extension of Resnik's association score, which considers dependencies of more than one case-markers in a subcategorization frame.

[7] Another related measure is Dunning (1993)'s likelihood ratio tests for binomial and multinomial distributions, which are claimed to be effective even with very much smaller volumes of text than is necessary for other tests based on assumed normal distributions.

[8] In the current implementation, we deal with sense ambiguities of case-marked nouns and case ambiguities of Japanese topic-marking post-positional particles such as *"ha(TOPIC)"*, *"mo(ALSO)"*, and *"dake(ONLY)"*. When constructing the set $F(e)$, we consider all the possible combination of senses of semantically ambiguous nouns and cases of topic-marking post-positional particles. These ambiguities can be resolved by maximizing the subcategorization preference function (section 5.2.2).

Table 1: The Result of Learning Probabilistic Subcategorization Preference for *"kau(buy,incur)"* (ϕ_{kl}, $\alpha = 0.9$)

	$\hat{F} = \{\langle f_{p_1}, \ldots, f_{p_n} \rangle\}$(Eg.)	$\hat{\phi}_{kl}$	Egs.
1	[*wo*(ACC):14(Products)]	1.88	158
2	[*wo*(ACC):13721-8(*kabu(stock)*)]	0.27	15
3	[*ga*(NOM):12(Human)]	0.27	40
4	[*wo*(ACC):15(Nature)]	0.21	25
5	[*kara*(from):12(Human)]	0.19	14
6	[*de*(at):12(Shop,Place)]	0.17	18
7	[*ga*(NOM):12(Human), *wo*(ACC):13721-8(*kabu(stock)*)]	0.16	6
8	[*wo*(ACC):13010(*hukyou(disgust)*)]	0.12	6
9	[*wo*(ACC):11961-1(Currency)]	0.10	6
10	[*ga*(NOM):12(Human),*wo*(ACC): 1456(Musical Instruments)]	0.09	4
	(11th~150th)	—	196

2. Iterative Reestimation

The subcategorization preference $\hat{\phi}(e)$ are maximized by repeatedly searching the set $F(e)$ for tuples $\langle f_1, \ldots, f_n \rangle$ which give the maximum subcategorization preference and removing other tuples from $F(e)$. The following two steps are repeated until the values of the parameters $q(f_k \mid v)$ converge.

(2a) For each verb-noun collocatoin e in $\mathcal{E}$, set $\hat{F}(e)$ as the set of tuples $\langle f_1, \ldots, f_n \rangle$ of independent partial subcategorization frames which can generate e and give the maximum subcategorization preference in the equation (23).

$$\hat{F}(e) \leftarrow \left\{ \langle f_1, \ldots, f_n \rangle \middle| \phi(\langle f_1, \ldots, f_n \rangle \rightarrow e) = \hat{\phi}(e) \right\}$$

(2b) Set the values of the conditional probabilities $p(\langle f_1, \ldots, f_n \rangle_j \mid e)$ as below and the parameters $q(f_k \mid v)$ as (17), respectively:

$$p(\langle f_1, \ldots, f_n \rangle_j \mid e) \leftarrow \frac{1}{|\hat{F}(e)|}$$

6 Experiments and Evaluation

6.1 Corpus and Thesaurus

As the training and test corpus, we used the EDR Japanese bracketed corpus (EDR, 1995), which contains about 210,000 sentences collected from newspaper and magazine articles. From the EDR corpus, we extracted 153,014 verb-noun collocations of 835 verbs which appear more than 50 times in the corpus. These verb-noun collocations contain about 270 case-markers. We constructed the training set $\mathcal{E}$ from these 153,014 verb-noun collocations.

We used 'Bunrui Goi Hyou'(BGH) (NLRI, 1993) as the Japanese thesaurus. BGH has a six-layered abstraction hierarchy and more than 60,000 words are assigned at the leaves and its nominal part contains about 45,000 words. Five classes are allocated at the next level from the root node.

6.2 Experiments and Results

From the training set $\mathcal{E}$, we first estimated the values of verb-independent parameters as in section 5.2.1, and then iteratively reestimated verb-dependent parameters of the subcategorization preference function ϕ_{kl} for 10 verbs as in section 5.2.2. For each of the 10 verbs, the numbers of verb-noun collocations are 100 ~ 500. We made experiments with the independence parameter $\alpha = 0.5/0.7/0.9$. In the iterative reestimation procedure, the values of the verb-dependent parameters converged after 2 ~ 5 iterations.

For the 10 verbs, about 75% of the verb-noun collocations have only one case-marked noun. The rate that tuples of partial subcategorization frames are judged as *independent* increases as the value of the independence parameter α decreases. This rate increases from 1.4% ($\alpha = 0.9$) to 12.1% ($\alpha = 0.5$).

As an example, for the verb *"kau(buy,incur)"*, Table 1 shows the set $\hat{F}(e)$ of tuples of independent partial subcategorization frames which give maximum subcategorization preference. The table lists the sets $\hat{F}(e)$ with 10 highest preference values of $\hat{\phi}_{kl}$, along with the numbers (the column 'Egs.') of verb-noun collocations for each $\hat{F}(e)$, which are judged as generated from it[9]. Since about 75% of the verb-noun collocations have only one case-marked noun, most of the 10 high-scored sets have only one case-marked noun. However, the 10 high-scored sets cover about 60% of the verb-noun collocations in the training set, and they can be regarded as typical subcategorization frames of the verb *"kau(buy,incur)"*.

6.3 Evaluation of Subcategorization Preference

We evaluate the performance of the estimated parameters of the subcategorization preference as follows.

Suppose that the following word sequence represents a verb-final Japanese sentence with a subordinate clause, where $N_x, \ldots, N_{2k}$ are nouns, $p_x, \ldots, p_{2k}$ are case-marking post-positional particles, v_1, v_2 are verbs, and the first verb v_1 is the head verb of the subordinate clause.

$$N_x\text{-}p_x\text{-}N_{11}\text{-}p_{11}\text{-}\cdots\text{-}N_{1l}\text{-}p_{1l}\text{-}v_1\text{-}N_{21}\text{-}p_{21}\text{-}\cdots\text{-}N_{2k}\text{-}p_{2k}\text{-}v_2$$

We consider the subcategorization ambiguity of the post-positional phrase N_x-p_x: i.e, whether N_x-p_x is subcategorized by v_1 or v_2.

We use held-out verb-noun collocations of the verbs v_1 and v_2 which are not used in the training. They are like those verb-noun collocations e_{c1} and e_{c2} in the left side below. Next, we generate erroneous verb-noun collocations e_{e1} of v_1 and e_{e2} of v_2 as those in the right side below, by choosing a case element $p_x : N_x$ at random and moving it from v_1 to v_2.

$$\begin{array}{lcl}
\boxed{\text{Correct}} & & \boxed{\text{Error}} \\
e_{c1} = \begin{bmatrix} pred : v_1 \\ p_{11} : N_{11} \\ \vdots \\ p_{1l} : N_{1l} \\ \underline{p_x : N_x} \end{bmatrix} & & e_{e1} = \begin{bmatrix} pred : v_1 \\ p_{11} : N_{11} \\ \vdots \\ p_{1l} : N_{1l} \end{bmatrix} \\
 & \Longleftrightarrow & \\
e_{c2} = \begin{bmatrix} pred : v_2 \\ p_{21} : N_{21} \\ \vdots \\ p_{2l} : N_{2k} \end{bmatrix} & & e_{e2} = \begin{bmatrix} pred : v_2 \\ p_{21} : N_{21} \\ \vdots \\ p_{2k} : N_{2k} \\ \underline{p_x : N_x} \end{bmatrix}
\end{array}$$

[9] In each subcategorization frame, Japanese noun classes of BGH thesaurus are represented as numerical codes, in which each digit denotes the choice of the branch in the thesaurus.

Table 2: Accuracies of Subcategorization Preference with ϕ_{kl} (%)

		Independent		Any	
		$\alpha=0.5$	$\alpha=0.9$	$\alpha=0.5$	$\alpha=0.9$
Optimal	+	81.7	70.7	65.8	68.6
	−	2.2	3.3	27.1	6.0
Initial	+	16.1	25.6	7.1	25.0
	−	0	0.4	0	0.4
Accuracy		97.8	96.3	72.9	93.6
Applicability		83.9	74.0	92.9	74.6

Then, we compare the sum $\hat{\phi}(e_{c1}) + \hat{\phi}(e_{c2})$ of the maximums (in the definition (23)) of ϕ_{kl} for the correct pair with the sum $\hat{\phi}(e_{e1}) + \hat{\phi}(e_{e2})$ of those for the erroneous pair, and calculate the rate that the correct pair has the greater value.

For the purpose of evaluating the effectiveness of factors of learning probabilistic subcategorization preference, we perform experiments with different settings and compare their results. The following two options are examined:

- Whether the subcategorization preference function uses tuples of partial subcategorization frames *judged as independent* ("Independent"), or *any* tuples ("Any").
- The independence parameter $\alpha=0.5/0.9$.

For three Japanese verbs *"kau (buy,incur)"*, *"nomu (drink)"*, and *"kasaneru (pile up, repeat)"*, we extracted pairs of correct verb-noun collocations and evaluated the performance of subcategorization preference. Table 2 gives the results averaged over extracted pairs, including the accuracies of subcategorization preference. The difference of "Optimal"/"Initial" means that initial values of the parameters are used instead of optimized values (section 5.2.2) when the subcategorization preference function is not applicable to the given verb-noun collocation and returns zero. The line "Accuracy" lists the sums of both "Optimal" and "Initial" accuracies, while the line "Applicability" lists the percentages of positive values of the subcategorization preference function with *optimized* parameters.

It is natural that the settings with more weak conditions on the independence judgment of partial subcategorization frames result in higher applicabilities. The setting with *independent* tuples of partial subcategorization frames achieves higher accuracy than that with *any* tuples, and this result claims that the result of the independence judgment is effective when applying the estimated parameters to the task of subcategorization preference. Even in the case of the setting with *any* tuples, the setting with $\alpha=0.5$ gives poorer accuracy than that of $\alpha=0.9$. In this case, the difference of the independence parameter α affects only the parameter estimation stage. This result claims that the independence judgment process is effective also when estimating parameters from the training corpus.

7 Conclusion

This paper proposed a novel method of learning probabilistic subcategorization preference of verbs. We described a part of the results of the experiments on learning probabilistic subcategorization preference from the EDR Japanese bracketed corpus, as well as those on evaluating the performance of subcategorization preference. Although the scale of the evaluation experiment was relatively small, we achieved accuracies higher than 96%. The details of the experimental results are available in Utsuro and Matsumoto (1997). As we mentioned in section 3.5, probabilistic model construction methods might be also applicable to the task of learning probabilistic subcategorization preference. We have already applied the maximum entropy methods(Pietra, Pietra, and Lafferty, 1995; Berger, Pietra, and Pietra, 1996) to this task(Utsuro, Miyata, and Matsumoto, 1997) and are also planning to evaluate the effectiveness of the MDL principle(Rissanen, 1989) when combining with the maximum entropy method. Their results will be compared with those of the method proposed in this paper and reported in the near future.

References

Baum, L. E. 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8.

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

Black, E. 1993. Towards history-based grammars: Using richer models for probabilistic parsing. In *Proceedings of the 31st Annual Meeting of ACL*, pages 31–37.

Collins, M. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of ACL*, pages 184–191.

Cover, T. M. and J. A. Thomas. 1991. *Elements of Information Theory*. John Wiley and Sons, Inc.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

EDR, (Japan Electronic Dictionary Research Institute, Ltd), 1995. *EDR Electronic Dictionary Technical Guide*.

Haruno, M. 1995. Verbal case frame acquisition as data compression. In *Proceedings of the 5th International Workshop on Natural Language Understanding and Logic Programming*.

Li, H. and N. Abe. 1995. Generalizing case frames using a thesaurus and the MDL principle. In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, pages 239–248.

Li, H. and N. Abe. 1996. Learning dependencies between case frame slots. In *Proceedings of the 16th COLING*, pages 10–15.

Magerman, D. M. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of ACL*, pages 276–283.

NLRI, (National Language Research Institute), 1993. *Word List by Semantic Principles*. Syuei Syuppan. (in Japanese).

Pietra, S. D., V. D. Pietra, and J. Lafferty. 1995. Inducing features of random fields. CMU Technical Report CMU-CS-95-144, School of Computer Science, Carnegie Mellon University.

Resnik, P. 1993. Semantic classes and syntactic ambiguity. In *Proceedings of the Human Language Technology Workshop*, pages 278–283.

Rissanen, J. 1989. *Stochastic Complexity in Statistical Inquiry*, volume 15 of *Series in Computer Science*. World Scientific Publishing Company.

Utsuro, T. and Y. Matsumoto. 1997. Learning probabilistic subcategorization preference and its application to syntactic disambiguation. Information Science Technical Report NAIST-IS-TR97006, Nara Institute of Science and Technology. (http://www.aist-nara.ac.jp/IS/TechReport/ report_gz/97006.ps.gz).

Utsuro, T., T. Miyata, and Y. Matsumoto. 1997. Maximum entropy parameter learning of subcategorization preference. (submitted to *the 35th Annual Meeting of ACL*).