

An Unsupervised Method for Canonicalization of Japanese Postpositions

Kentaro Torisawa

Graduate School of Information Sciences,

Japan Advanced Institute of Science and Technology

1-1 Asahidai, Tatsunokuchi-machi, Nomi-gun, Ishikawa, 923-1292 JAPAN

Information and Human Behavior, PRESTO, Japan Science and Technology Corporation

4-1-8 Kawaguchi Hon-cho, Kawaguchi-shi, Saitama, 332-0012 JAPAN

torisawa@jaist.ac.jp

Abstract

We present an unsupervised method for canonicalizing *joshi* (postpositions) in Japanese. Some postpositions in Japanese do not specify semantic roles explicitly as case markers do, although those postpositions syntactically behave as the case markers. Such postpositions includes “wa,” which *topicalizes* noun phrases, and “mo,” which *emphasizes* noun phrases. For this paper, we replaced these postpositions in a sentence with case markers, without changing the meanings of the original sentence as little as possible. This leads to canonicalization or paraphrasing of verb phrases into canonical forms with desirable properties. Our method utilized case frames and semantic word classifications induced by the Expectation Maximization algorithm. The induction process was unsupervised in the sense that no semantic clues were given before the induction of the case frames and the word classifications.

1 Introduction

We developed an unsupervised method to canonicalize postpositions in Japanese. Some postpositions in Japanese do not specify semantic roles explicitly as case markers do, although they syntactically behave as the case markers. These postpositions include “wa,” which *topicalizes* noun phrases, and “mo,” which *emphasizes* noun phrases. We call such postpositions *non case marking postpositions* and distinguish them from case markers such as “wo” and “ga.” In this paper, we replaced the non case marking postpositions with case markers while retaining the original meaning as well as possible. We called this process *postposition canonicalization*. It leads to canonicalization or paraphrasing of verb phrases into canonical forms with the following desirable properties.

- Semantic roles of noun phrases can be determined easily.

- Semantically valid matching between canonical forms can be easily done.

The first property can contribute to improvement of rather traditional NLP systems, which translate natural language inputs to formal expressions such as logical forms. Semantic roles must be determined during this translation process. If we can replace non case marking postpositions with case markers, the determination of semantic role becomes an easier task and accuracy can be improved. On the other hand, the second property of canonical form matching may contribute to improvement of a rather new type of NLP systems. These systems do not utilize artificial semantic representations during processing. They process natural language on a level close to a surface form of the language. There are attempts to perform even inferences about natural language inputs on a shallow level by performing direct matching between parse trees of the natural language inputs (Kurohashi and Higasa, 2000). Even in these systems, the existence of non case marking postpositions is harmful because a sentence with such postpositions can be seen as an *ambiguous* sentence with respect to semantic roles. Then, non case marking postpositions can cause incorrect and unintended matching between shallow level expressions. The postposition canonicalization may help to reduce this ambiguity and contribute to improvement accuracy in such a system.

Our method utilized case frames and semantic word classifications produced by the Expectation Maximization algorithm (EM algorithm) (Dempster et al., 1977) from parsing results of news paper articles. The method was unsupervised in the sense that we did not give any semantic clues for constructing case frames and semantic word classifications. The semantic frames and word classification were expressed by probabilistic distributions. In postposition canonicalization, we began with a triple, consisting of a head verb, a non case-marking postposition and a noun. Then, we determined a case marker with which the given noun-verb pair cooccurred with the highest likelihood, and replace the postpositions with the selected case markers.

Kawahara et al., attempted canonicalization of Japanese postpositions by using manually tailored thesaurus (Kawahara et al., 2000). A fair comparison between Kawahara’s method and ours is difficult because of the difference in experimental settings, we show that our results are comparable, although we used automatically learned word classifications instead of hand crafted thesauri.

This paper is organized as follows. Section 2 overviews Japanese postpositions. Section 3 gives a brief description of EM-based unsupervised learning. Section 4 describes an algorithm for postposition canonicalization. Section 5 shows a series of experimental results. Finally, Section 6 outlines future works and presents concluding remarks.

2 Japanese Postpositions

Japanese is a head-final language, which allows rather free word order scrambling. In such a language, word order is not used for encoding semantic roles. Instead, postpositions play an important role in determining the semantic roles of noun phrases in sentences. Most postpositions behave as case markers, which specify semantic roles. For instance, the two sentences below mean “Taro writes a paper” although they have different word orders. The postpositions “ga” and “wo” work as case markers. “ga” specifies an agent semantic role and “wo” a theme role. Note that “Taro” is a person’s name, and that a case marker modifies the noun to its left.

- Taro ga ronbun wo kaku.
 noun noun verb
 Taro a paper write
- ronbun wo Taro ga kaku.
 a paper Taro write

Case markers have a certain degree of ambiguities in specifying semantic roles. With some verbs, the case marker “ga” can specify the theme role, not the agent role as in the above example. However, in most cases, we cannot replace case markers in a sentence without changing the meaning of the original sentence. The following sentences, which were obtained by replacing the case markers in the above examples, do not preserve the same meaning as the above ones. We did not give English translations to syntactically unacceptable sentences.

- *Taro wo ronbun ga kaku.
 Taro a paper write
 (A paper writes Taro.)
- *Taro ga ronbun ga kaku.
 Taro a paper write
- *Taro wo ronbun wo kaku.
 Taro a paper write

However, there are some postpositions which can replace case markers rather freely. The following sentences have the same meaning as the first sentences, except for topicalization and emphasis caused by the newly inserted postpositions “wa” and “mo.”

- Taro wa ronbun wo kaku.
 Taro a paper write
- Taro ga ronbun wa kaku.
 Taro a paper write
- Taro mo ronbun wo kaku.
 Taro a paper write
- Taro ga ronbun mo kaku.
 Taro a paper write

The above examples suggest that the postpositions such as “wa” and “mo” cannot specify semantic roles explicitly as case markers do. This implies that those postpositions that erase case markers make it difficult, in a sense, to interpret the sentence. Another important point is that both of “wa” and “mo” are used quite frequently in daily life and that they are not the exceptional cases of interests only to linguists. In addition, there are extreme cases such as the following.

- ronbun wo kyou mo kaku
 a paper today write
 ((Someone) writes a paper today.)
- kyou wa ronbun wo kaku
 today a paper write
 ((Someone) writes a paper today.)
- ronbun wo kyou kaku
 a paper today write
 ((Someone) writes a paper today.)
- ronbun wo 2-hon mo kaku
 paper two write
 ((Someone) writes two papers.)
- ronbun wo 2-hon wa kaku
 paper two write
 ((Someone) writes two papers.)
- ronbun wo 2-hon kaku
 paper two write
 ((Someone) writes two papers.)

In these cases, the postpositions do not have the corresponding case markers. The nouns modified by “wa” and “mo” can appear without any postpositions. Although “2-hon” semantically behaves as a quantifier that modifies the semantic contents of the noun phrase “ronbun”, it, syntactically, is a modifier of the head verb “kaku.” This phenomenon is called “quantifier floating.” Although proper linguistic analysis of these phenomena would be quite complex, in this paper we regard these as situations where a noun phrase is modified by an *empty* case marker.

We call the postpositions such as “wa” and “mo” *non case marking postpositions*, and distinguish them from the case markers such as “ga” and “wo.” As mentioned before, non case-marking postpositions do not specify semantic roles as case markers do. As a result, semantic interpretation of sentences with non case-marking postpositions can be more difficult than that of the corresponding sentences that include only case markers.

The task addressed in this paper was to replace non case-marking postpositions, such as “wa” and “mo,” with case markers, such as “ga” and “wo,” or an *empty* case marker, without changing the meaning of sentences. We called this process *postposition canonicalization*. The outcome of this process was sentences that included only case markers, and it was easier to determine proper semantic roles for noun phrases in such sentences than in the original sentences with non case-marking postpositions.

Our basic approach for the postposition canonicalization was to utilize the semantic case frames and word classifications that were induced by an unsupervised learning method. By using the case frames and word classifications, we could estimate the likelihood that a case marker replaced a given non case marking postposition. Note that, in this work, the semantic case frames and word classifications were represented as a probabilistic distribution, and that this enabled us to estimate the likelihood directly. Then, according to the estimated likelihood, we determined the case marker that should replace the given case marking postposition.

3 An EM-based Unsupervised Learning Method

We utilized the Expectation Maximization (EM) algorithm for learning semantic case frames and word classifications, which are required for the postposition canonicalization. We followed the methods presented in Rooth, et al., 1999; and Hofmann and Puzicha, 1998, basically in realizing learning methods, and also added some extensions.

The method required triples in the form of $\langle v, rel, n \rangle$ as learning data, where v was a verb, n was a head noun, and rel stood for the relationship between v and n . In our study, rel was called *cooccurrence relation* and could be one of the postpositions, including case markers and non case marking postpositions. In addition, we assumed that the empty case marker mentioned in the previous section could be the value of rel .

Each triple $\langle v, rel, n \rangle$, which we called *cooccurrence triple*, was divided into two items. The first was $\langle v, rel \rangle$ and the second n . we called $\langle v, rel \rangle$ the *argument position*. Then, the probability that the cooccurrence triple $\langle v, rel, n \rangle$ appeared was defined as follows.

$$P(\langle v, rel, n \rangle) =_{def} \sum_{a \in A} P(\langle v, rel \rangle | a) P(n | a) P(a)$$

where a denoted a *class* of the cooccurrences. More intuitively, it corresponded to a semantic class of a noun n and an argument position denoted by $\langle v, rel \rangle$. Here, we assumed that A is a set of k integers denoted by $\{1, 2, 3, \dots, k\}$, where the integer k was given before learning.

What the EM-based clustering method does is to estimate each probabilities denoted by $P(\langle v, rel \rangle | a)$, $P(n | a)$, $P(a)$ for each verb v , noun n , cooccurrence relation rel , and class a . Here, we can regard the probability $P(n | a)$ as a representation of a word classification. We can derive the probability $P(a | n)$ from $P(n | a)$ by applying the Bayes’ Theorem. $P(a | n)$ can be interpreted as the probability that n ’s appearance is used as a word in the class a . Note that we can compute all $P(a | n)$ for each class denoted by a , and that this distribution can be seen as the distribution of distinct usages of the word n . This means that the classification expressed by $P(a | n)$ can capture classifications and ambiguities of the usages of the word n . In the same way, we can derive the probability $P(a | \langle v, rel \rangle)$ from $P(\langle v, rel \rangle | a)$, and can regard $P(a | \langle v, rel \rangle)$ as a classification of usages of the argument position $\langle v, rel \rangle$. In the remainder of the paper, we regard $P(\langle v, rel \rangle | a)$ and $P(a | \langle v, rel \rangle)$ for a given verb v and a cooccurrence relation rel as a case frame of the verb v .

Unfortunately, estimation of the probabilities $P(\langle v, rel \rangle | a)$, $P(n | a)$ and $P(a)$ is not straightforward, since the class a is not observed in a given corpus. The EM-based clustering method estimates these probabilities with the unobserved data a by the following iterative procedure.

First, we consider the list of cooccurrence relations observed in a given corpus, which is actually a set of parse trees obtained by a statistical parser. Let us create the list

$$L = \langle \langle v_0, rel_0, n_0 \rangle, \langle v_1, rel_1, n_1 \rangle, \dots, \langle v_m, rel_m, n_m \rangle \rangle.$$

Then, the likelihood that L is observed is denoted by the following formula.

$$\prod_{\langle v_i, rel_i, n_i \rangle \in L} P(\langle v, rel, n \rangle) = \prod_{\langle v_i, rel_i, n_i \rangle \in L} \left\{ \sum_{a \in A} P(\langle v_i, rel_i \rangle | a) P(n_i | a) P(a) \right\}$$

The EM algorithm is an instance of a maximum likelihood estimation and it maximizes the above likelihood by adjusting the parameters $\{P(\langle v, rel \rangle | a) | v \in V, rel \in Rel, a \in A\} \cup \{P(n | a) | n \in N, a \in A\} \cup \{P(a) | a \in A\}$ in an iterative fashion, where V denotes the set of all the observed verbs, N a set of all the observed nouns, and Rel is a set of all the possible relations. The iteration continues until the likelihood converges

at a certain value, or iteration steps of a given number are completed.

We denote the probabilities computed at the j -th iteration step by $P_j(\cdot)$. The initial values of the probabilities are denoted by $P_0(\cdot)$. Although the initial probabilities were often determined randomly in previous works, we developed a method to determine the initial values so that the convergence of the likelihood was accelerated. We describe the method later in this section.

The probabilities $P_{j+1}(\cdot)$ are computed from $P_j(\cdot)$ in the following manner. Assume that the procedure has just finished the j -th iteration and has computed the probabilities denoted by $P_j(\cdot)$. First, we need to compute $P_j(a|\langle v, rel \rangle, n)$ by the following formula.

$$P_j(a|\langle v, rel \rangle, n) = \frac{P_j(a)P_j(\langle v, rel \rangle|a)P_j(n|a)}{\sum_{a' \in A} P_j(a')P_j(\langle v, rel \rangle|a')P_j(n|a')}$$

Then, the new probabilities to be computed at the $j + 1$ -th iteration are given in the following formulae.

- $P_{j+1}(a) = \frac{1}{m} \sum_{\langle v_i, rel_i, n_i \rangle \in L} P_j(a|\langle v, rel \rangle, n)$
- $P_{j+1}(\langle v, rel \rangle|a) = \frac{\sum_{\langle v, rel, n_i \rangle \in L} P_j(a|\langle v, rel \rangle, n_i)}{\sum_{\langle v_i, rel_i, n_i \rangle \in L} P_j(a|\langle v_i, rel_i \rangle, n_i)}$
- $P_{j+1}(n|a) = \frac{\sum_{\langle v_i, rel_i, n \rangle \in L} P_j(a|\langle v_i, rel_i \rangle, n)}{\sum_{\langle v_i, rel_i, n_i \rangle \in L} P_j(a|\langle v_i, rel_i \rangle, n_i)}$

These functions are derived by applying standard derivation steps in the EM method. In short, the probabilities are adjusted so that they reflect the probabilities of observed data more closely as the iterations proceed. If the process terminates at the m -th iteration, the probabilities denoted by $P_m(\cdot)$ are the outcome of the entire learning process. It is empirically known that the probabilistic distribution obtained reflects semantic classifications and case frames for English. This was also confirmed for Japanese as described in Section 5.

Actually, we extended the above method in the following two respects.¹

1. Preprocessing by an clustering method with average mutual information.
2. Utilization of the cooccurrence likelihood obtained by a statistical parser.

Let us give a brief overview of the extensions. At the beginning of the EM-based learning algorithm, one must give the initial probabilities denoted by $P_0(\cdot)$. Although these values can

¹In the experiments presented in this paper, we used a statistical model extended in another aspect. The extension aimed at treating not only noun-verb relationship, but also noun-noun-verb relationships in probability distributions. See (Torisawa, 2001) for the details of the extended model.

be determined randomly, we attempted another method. In our method, the nouns were classified by applying the clustering method with average mutual information (Brown et al., 1992) to the cooccurrence triples. This method could give us semantically acceptable classifications of words, although the classification were mutually exclusive and it did not allow any ambiguity in the classification. Then, the initial probabilities were determined according to the classification obtained. More precisely, if a word w belonged to the class a , then we set $P(a|w) = P$ and $P(a'|w) = (1 - P)/k$ for the class a' other than a , where P was a fixed, high probability and k the number of the classes. The initial probabilities $P_0(\cdot)$ could be computed from those probabilities. We found that this process sped up convergence of the EM-based learning method.

In addition, we augmented the cooccurrence triples with the cooccurrence likelihoods obtained by a statistical parser. The statistical parser computed the likelihood of the cooccurrence triples during parsing. We weighted the frequencies of the given cooccurrence triples with such likelihoods. In general, if the likelihood of a cooccurrence triple was small, then the cooccurrence triple was likely to be erroneously produced by parsing errors. By this method, the effects caused by parsing errors could be reduced, although we have not confirmed this by experiments.

4 Postposition Canonicalization Algorithm

Given the semantic case frames and the word classifications induced by the EM algorithm, one can determine the case markers that replace given non case marking postpositions such as “wa” and “mo.” Consider the situation in the following sentences, where n is a noun and v is a verb.

- $n \text{ wa } v$
- $n \text{ mo } v$

By using the probabilistic distribution obtained by the EM algorithm, one can compute the probabilities of the cooccurrences including n , v and a case marker such as “ga” or “wo.” The probability of the cooccurrence is given by the following formula, where cm stands for a case marker. Note that that the probabilities $P(\langle v, cm \rangle|a)$, $P(n|a)$ and $P(a)$ are estimated by the EM algorithm presented in the previous section.

$$P(\langle v, cm, n \rangle) = \sum_{a \in A} P(\langle v, cm \rangle|a)P(n|a)P(a)$$

The case marker cm that should replace the non case marking postpositions is given by the following formula. Basically, it selects a case marker according to the probability $P(\langle v, cm, n \rangle)$ obtained

by the above formula. Let us assume that C denotes the set of all the case marker that can replace non case marking postpositions.

$$CM(v, n) = \operatorname{argmax}_{cm \in C} W(cm)P(\langle v, cm, n \rangle)$$

$W(cm)$ is a weight factor reflecting preferences on a case marker cm . The likelihood that a given non case marking postposition is replaced with a case marker depends not only on $P(\langle v, cm, n \rangle)$ but also on the non case marking postposition to be replaced. For instance, “mo” can be replaced with case makers with essentially uniform likelihoods. But “wa” has a strong preference for the case marker “ga.” In the above formula, $W(cm)$ is used for expressing such preferences. In the experiments described in the next section, we used a uniform weight $W(cm) = 1.0$ for every $cm \in C$ in the canonicalization of “mo.” But, in the canonicalization of “wa,” we set $W(ga) = 30.0$ and $W(cm) = 1.0$ for other case markers cm in $C - \{ga\}$.

In the experiments described in the next section, we assumed $C = \{ga, wo, e\}$ where e denoted the *empty* case marker. Although there can be the case where non case marking postpositions are replaced with other case markers such as “ni” or “de,” such situation occurs rather infrequently, and we assumed that those case markers were negligible in our task.

In addition, there are some cases that the probability $P(\langle v, cm \rangle | a)$ was not estimated because the argument position $\langle v, cm \rangle$ was observed less frequently than the threshold we set in the experiments. This suggested that the position $\langle v, cm \rangle$ is linguistically ill-formed. But there were some v such that $\langle v, ga \rangle$ was observed with only a small frequency. Since all the Japanese verbs can, in theory, take the case marker “ga” as its argument, we cannot regard this as a suggestion that the argument position does not exist. In such a case, our algorithm provided “ga” as the case marker that replaced a given postposition, without considering other case markers. This is justified by the fact that the most frequently replaced case marker was “ga.”

Then, the whole algorithm is summarized as the following re-formulated formula.

$$CM(v, n) = \begin{cases} ga & \text{if } P(\langle v, ga \rangle) \text{ is not estimated.} \\ \operatorname{argmax}_{cm \in C} W(cm)P(\langle v, cm, n \rangle) & \text{otherwise} \end{cases}$$

By replacing non case marking postpositions with $CM(v, n)$, one can convert the verb phrases including non case marking postpositions into a canonical form that has only case markers.

5 Experiments

5.1 Clustering

- CLASS 43

- word	classifications
日立製作所 (HITACHI)	0.597
日本ビクター (JVC)	0.558
シャープ (SHARP)	0.513
三菱電機 (MITSUBISHI)	0.502
東芝 (TOHSHIBA)	0.493
NEC (NEC)	0.481
三洋電機 (SANYO)	0.481
富士通 (FUJITSU)	0.468
アイワ (AIWA)	0.465
川崎重工業 (KAWASAKI)	0.447
リコー (RICOH)	0.414
ソニー (SONY)	0.410
沖電気工業 (OKI)	0.408

- case	frames
\langle 三菱電機, $e \rangle$ (\langle MITSUBISHI, $e \rangle$)	0.456
\langle 日立製作所, $e \rangle$ (\langle HITACHI, $e \rangle$)	0.419
\langle 東芝, $e \rangle$ (\langle TOHSHIBA, $e \rangle$)	0.415
\langle 富士通, $e \rangle$ (\langle FUJITSU, $e \rangle$)	0.405
\langle 三洋電機, $e \rangle$ (\langle SANYO, $e \rangle$)	0.352
\langle シャープ, $e \rangle$ (\langle SHARP, $e \rangle$)	0.350
\langle キヤノン, $e \rangle$ (\langle CANON, $e \rangle$)	0.338
\langle NEC, $e \rangle$ (\langle NECT, $e \rangle$)	0.337
\langle 発売する, $ga \rangle$	
\langle (\langle hatsubai' suru, $ga \rangle$)	
\langle (\langle (start to sell), $ga \rangle$)	0.280

- CLASS 869

- word	classifications
エア-ニッポン (Air Nippon)	0.562
エアシステム (Air System)	0.448
ANK (ANK)	0.423
航空 (airline)	0.412
空輸 (air cargo)	0.370
全日空 (All Nippon Airline)	0.354
汽船 (steam ship)	0.323
日航 (Japan Airline)	0.292
スカイマークエアラインズ (Skymark Airlines)	0.28597
エア (airline)	0.280
JAS (Japan Air System)	0.246
海運 (cargo)	0.234
JAL (Japan Airlines)	0.232
ノースウエスト (Northwest)	0.199
ユナイテッド (United)	0.198

- case	frames
\langle エアシステム, $e \rangle$ (\langle Air System, $e \rangle$)	0.340
\langle 空輸, $e \rangle$ (\langle Kuuyu (air cargo), $e \rangle$)	0.330
\langle 全日空, $e \rangle$ (\langle All Nippon Airline, $e \rangle$)	0.310
\langle 運航する, $ga \rangle$	
\langle (\langle unkou' suru, $ga \rangle$)	
\langle (\langle (run cargos or airlines), $ga \rangle$)	0.302

- CLASS 1588

- word	classifications
大阪 (Osaka, a city)	0.477
名古屋 (Nagoya, a city)	0.215
仙台 (Sendai, a city)	0.215
札幌 (Sapporo, a city)	0.164

- case	frames
\langle 名古屋, $e \rangle$ (\langle Nagoya (a city), $e \rangle$)	0.373
\langle 京都, $e \rangle$ (\langle Kyoto (a city), $e \rangle$)	0.310
\langle 福岡, $e \rangle$ (\langle Fukuoka (a city), $e \rangle$)	0.251

Figure 1: Clusters in the clustering results. (when the number of the classes is 2000.)

- CLASS 1730
 - word classifications

酒 (drink with alcohol)	0.486
お茶 (green tea)	0.444
コーヒー (coffee)	0.331
日本酒 (Japanese sake)	0.317
ビール (beer)	0.288
ウーロン茶 (oolong tea)	0.282
シャンパン (champagne)	0.276
紅茶 (tea)	0.273
 - case frames

〈 飲む, を 〉	0.304
〈 (<i>nomu</i> (drink), <i>wo</i>) 〉	
〈 酔う, に 〉	0.206
〈 (<i>you</i> (be drunken), <i>ni</i>) 〉	
- CLASS 165
 - word classifications

スナック (bar)	0.048
居酒屋 (pub)	0.045
喫茶店 (cafe)	0.034
飲食店 (restaurant)	0.032
食堂 (restaurant)	0.031
料亭 (restaurant)	0.030
 - case frames

〈 飲む, で 〉	0.113
〈 (<i>nomu</i> (drink), <i>de</i>) 〉	
〈 移転する, から 〉	0.012
〈 (<i>iten'suru</i> (move), <i>kara</i>) 〉	
- CLASS 1143
 - word classifications

一橋 (Hitotsubashi Univ.)	0.908
立命館 (Ritsumeikan Univ.)	0.788
慶 (Keio Univ.)	0.649
筑波 (Tsukuba Univ.)	0.593
早稲田 (Waseda Univ.)	0.550
 - case frames

〈 工学部, e 〉	0.337
〈 (<i>kogakubu</i> (dept. of engineering), <i>e</i>) 〉	
〈 教授, e 〉	0.317
〈 (<i>kogakubu</i> (professor), <i>e</i>) 〉	
- CLASS 781
 - word classifications

ドジャース (Dodgers)	0.652
ブルズ (Bulls)	0.600
メッツ (Mets)	0.549
法大 (Hosei Univ.)	0.540
レッズ (Reds)	0.495
カーディナルス (Cardinals)	0.493
 - case frames

〈 連敗, は 〉	0.294
〈 (<i>renpai</i> (lose several games), <i>wa</i>) 〉	
〈 投手, の 〉	0.277
〈 (<i>toushu</i> (a pitcher), <i>no</i>) 〉	
〈 大宮 〉	0.263
〈 (<i>Ohmiya</i> (the name of a city), <i>e</i>) 〉	

Figure 2: Clusters in the clustering results.(Continued from Figure 1.)

- 早大 (Soudai)
 - $P(781 | \text{早大 (Soudai)}) = 0.20$
 - $P(1143 | \text{早大 (Soudai)}) = 0.09$
- 磐田 (Iwata)
 - $P(781 | \text{磐田 (Iwata)}) = 0.38$
 - $P(1588 | \text{磐田 (Iwata)}) = 0.02$

Figure 3: The ambiguities captured by the clustering results.

We obtained cooccurrence triples from the parsing results of 14 years of news paper articles. (9 years of Nikkei Shinbun and 5 years of Mainichi Shinbun) We used a down graded version of a hybrid parsing system utilizing both of a grammar and a statistical framework (Kanayama et al., 2000) in order to parse the corpora.

We also restricted the cooccurrence triples to those consisting of the words and argument positions that appeared more than thirty times in one year of news paper articles. As a result, we obtained 18,360 words and 25,473 argument positions.² Unlike the explanation in the previous section, the argument positions in the experiments included a modifier of nouns. For instance, if there was a NP modifying another NP as in the example sentence,

- NP_1 no NP_2 , (no is a postposition whose direct translation is “of” in English.)

we added a triple $\langle n_2, no, n_1 \rangle$ to the cooccurrence triples, where n_1 and n_2 are head nouns of NP_1 and NP_2 respectively. In a similar manner, we treated coordination of nouns and noun sequences in a complex noun in cooccurrence relations. If there is a noun sequence as the following example,

- NP_1 NP_2

we added the triple $\langle n_2, e, n_1 \rangle$ where e stands for the *empty* case marker and n_1 and n_2 are head nouns of NP_1 and NP_2 respectively. The total frequencies of cooccurrence triples weighted by dependency likelihood was 1.08×10^8 .³

In addition, when a verb was modified by certain auxiliaries in the corpus, we treated them as a single word. An example of such auxiliaries is “reru,” which passivizes the verb. We found that ignoring the auxiliary is harmful for clustering

²As mentioned before, we used an EM-based learning algorithm that can treat noun-noun-verb relationships as learning data. To do this, we also obtained 19,704 verb phrase template, which should be regarded as pairs of argument positions with common verbs. See Torisawa 2001 for more details.

³Again, we collected the triples each of which consists of two nouns and a verb phrase template from the same corpus. The total frequency of such triples was 1.61×10^7 .

# of classes	# of correct case markers	accuracy (%)
BASELINE	467	69.5
500	556	82.7
1000	565	84.1
2000	567	84.4

Figure 4: The results of the postposition canonicalization for the postposition “mo.”

results since this auxiliary changes the semantic roles of noun phrase.

We applied the EM algorithm to the data given three different numbers of classes, namely, 500, 1000 and 2000. The probabilities estimated by the EM algorithm was obtained by 40 iterations passes for each number of classes. Some of the classes obtained are presented in Figure 1 and Figure 2. The tables give the argument positions with the highest $P(a|\langle v, rel \rangle)$ and the nouns with the highest $P(a|n)$ with those probabilities. These probabilities were derived from the estimated parameters $P(\langle v, rel \rangle|a)$, $P(n|a)$ and $P(a)$ by applying the Bayes’ theorem. We found that the results were intuitively acceptable. The classes, other than those presented in the figures, had the semantic uniformity, which could be recognized easily.

The most remarkable point was that the clustering results could capture some ambiguities that were unlikely to appear in a manually tailored lexicon. Figure 3 gives some examples. Note that the probabilities given in this figure are taken from the same distribution as in Figure 1 and Figure 2. The integers identifying the classes were common to all the figures.

“Soudai” is the name of a university, but it is often used as the name of the sport teams from the university. The phenomenon is highly metonymical, and is unlikely to appear in normal dictionaries. But the estimated probabilities captured this ambiguity, as shown in the table. “Soudai” is used as an instance of the class 781, which included sport teams, with the probability 0.20, and appears with the probability 0.09 as a member of the class 1143, which is the class of universities. “Iwata” is a similar example. It refers to the name of a city, which should be categorized into class 1588. But the city has a well-known football club and this was reflected by the probability $P(781|Iwata)$, where the class 781 included sport teams.

5.2 Postposition Canonicalization

We extracted 794 cooccurrence triples, including the non case marking postposition “mo”, and 903 triples that included “wa” in the EDR bracketed corpus (EDR, 1996). Note that the coc-

# of classes	# of correct case markers	accuracy (%)
BASELINE	619	86.3
500	624	87.0
1000	632	88.1
2000	627	87.4

Figure 5: The results of the postposition canonicalization for the postposition “wa.”

currence triples were guaranteed to be correct, unlike the triples fed to the EM algorithm. In those triples with “mo,” 672 triples consisted of the words fed to the EM algorithm. In other words, 122 (= 794 – 672) triples included the unknown words for the learning algorithm. On the other hands, 717 triples were covered by our method in the case of “wa.” Then, the coverages of our current implementation were 84.6% for “mo” and 79.4% for “wa.” Note that the probabilities for some argument positions, which were pairs of words and cooccurrence relations, may not have been estimated, even if the words are covered by the learning data.

To each triple, we assigned a case marker that should replace the non case-marking postpositions, and compared those *correct* case markers with the results of our postposition canonicalization method. Although our algorithm produced only three case makers, namely “ga”, “wo” and “e”, where *e* was an *empty* case marker, our tagging scheme allowed other case markers, namely, “ni,” and “de.” But their frequencies were quite small and we assumed they were negligible. Japanese native speakers might think that there are many sentences that include time expressions with non case-marking postpositions, and the “ni” should replace such postpositions. But such occurrences of the postpositions can be replaced with the empty case marker “e” as well. In news paper articles, time expressions often appear without any case marker, i.e., with the empty case marker. Thus, we decided to employ “e” instead of “ni” as much as possible in assigning case markers.

We used three different probability distributions induced by the EM algorithm in postposition canonicalization. They have different numbers of classes, namely 500, 1000, and 2000. The results are presented in Figures 4 and 5. In each table, BASELINE refers to the performance obtained by assigning “ga” to all the cooccurrence triples. Note that “ga” replaced both of “wa” and “mo” most frequently.

The accuracies presented in the tables are the ratios of the triples that are assigned correct case markers to the number of the cooccurrence triples that did not include unknown words. Our method achieved more than 80% accuracy for both of “wa”

and “mo.” It contributed to a nearly 15% accuracy improvement compared to the baseline method in the canonicalization of “mo.” On the other hands, the improvement in the case of “wa” was less than 2%. Note that, in the canonicalization of “wa,” the baseline method already provided rather good results. If we apply the baseline method for the cooccurrence triples that included unknown words, the accuracies for all the extracted cooccurrence triples were expected to be 82.1% for “mo” by using the 2000 classes and 87.7% for “wa” by using the 1000 classes. These figures can be easily derived from the coverage of our method.

Kawahara, et al. reported about 82% accuracies for the mixture of “wa,” “mo,” and other *topic markers*. (Kawahara et al., 2000) Considering the distributions of those words in the corpus, our results are comparable to theirs although a fair comparison is difficult because of the difference of experimental settings. (For instance, their test set included parsing errors, while ours did not.) Besides a simple comparison of the accuracies, we think that our results are significant in the respect that we did not employ any manually tailored thesauri, as Kawahara did.

The number of classes affected the performance to a certain degree. The results suggested that the 500 class classification was too coarse classification for this task.

6 Future Work and Concluding Remarks

We have presented an unsupervised method for replacing the postpositions that do not specify semantic roles explicitly, with case markers that do specify semantic roles. We called this replacement *postposition canonicalization*. Our method utilized the case frames and word classifications learned by the Expectation Maximization algorithm from corpora. Although the method was rather simple, it achieved more than 80% accuracy for the postpositions “wa” and “mo.” We think that there are several possible extensions that can improve the performance. First of all, our method determined a case marker only by looking at a noun and a verb. But there were a situation when a guessed case marker had already been used in a given verb phrase. The case marker was unlikely to be the correct one in such a situation. Then, it was worth considering that other postpositions in the same verb phrases might contribute to higher accuracy. In addition, we considered only the case alternations caused by a few auxiliaries. The algorithm should be extended so that it can treat more case alternations caused by a wide range of auxiliaries. A more fundamental problem is how to capture the bias due to the postpositions to be replaced. We weighted the estimated probabilities according to human intuition in canonicalization of “wa.” It would be interesting to investigate how

to objectively compute such bias.

Besides the improvement of postposition canonicalization, there are more applications to which the EM-based clustering method can be applied. For instance, we expect that gap filling for relative clauses (Baldwin et al., 1999) can be performed in a method similar to the framework presented in this paper.

References

- Timothy Baldwin, Takenobu Tokunaga, and Hozumi Tanaka. 1999. The parameter-based analysis of Japanese relative clause constructions. In *IPSJ SIJ Notes*, volume 99(95), pages 55–62.
- Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jennifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):31–40.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38.
- EDR. 1996. *EDR electronic dictionary version 1.5 technical guide*.
- Hiroshi Kanayama, Kentaro Torisawa, Yutaka Mitsuishi, and Jun’ichi Tsujii. 2000. A hybrid Japanese parser with hand-crafted grammar and statistics. In *Proceedings of COLING 2000*, pages 411–417.
- Daisuke Kawahara, Nobuhiro Kaji, and Sadao Kurohashi. 2000. Japanese case structure analysis by unsupervised construction of a case frame dictionary. In *Proceedings of 18th International Conference on Computational Linguistics*, pages 432–438.
- Sadao Kurohashi and Wataru Higasa. 2000. Dialogue helpsystem based on flexible matching of user query with natural language knowledge base. In *Proceedings of 1st ACL SIGdial Workshop on Discourse and Dialogue*, pages 141–149.
- Kentaro Torisawa. 2001. A nearly unsupervised learning method for automatic paraphrasing of Japanese noun phrases. In *Proceedings of Workshop on Automatic Paraphrasing: Theories and Applications*, Tokyo, Japan. to appear.