Learning to Predict Case Markers in Japanese

Hisami Suzuki **Kristina Toutanova**[1]
Microsoft Research
One Microsoft Way, Redmond WA 98052 USA
{hisamis,kristout}@microsoft.com

Abstract

Japanese case markers, which indicate the grammatical relation of the complement NP to the predicate, often pose challenges to the generation of Japanese text, be it done by a foreign language learner, or by a machine translation (MT) system. In this paper, we describe the task of predicting Japanese case markers and propose machine learning methods for solving it in two settings: (i) *monolingual*, when given information only from the Japanese sentence; and (ii) *bilingual*, when also given information from a corresponding English source sentence in an MT context. We formulate the task after the well-studied task of English semantic role labelling, and explore features from a syntactic dependency structure of the sentence. For the monolingual task, we evaluated our models on the Kyoto Corpus and achieved over 84% accuracy in assigning correct case markers for each phrase. For the bilingual task, we achieved an accuracy of 92% per phrase using a bilingual dataset from a technical domain. We show that in both settings, features that exploit dependency information, whether derived from gold-standard annotations or automatically assigned, contribute significantly to the prediction of case markers.

1 Introduction: why predict case?

Generation of grammatical elements such as inflectional endings and case markers has become an important component technology, particularly in the context of machine translation (MT). In an English-to-Japanese MT system, for example, Japanese case markers, which indicate grammatical relations (e.g., subject, object, location) of the complement noun phrase to the predicate, are among the most difficult to generate appropriately. This is because the case markers often do not correspond to any word in the source language as many grammatical relations are expressed via word order in English. It is also difficult because the mapping between the case markers and the grammatical relations they express is very complex. For the same reasons, generation of case markers is challenging to foreign language learners. This difficulty in generation, however, does not mean the choice of case markers is insignificant: when a generated sentence contains mistakes in grammatical elements, they often lead to severe unintelligibility, sometimes resulting in a different semantic interpretation from the intended one. Therefore, having a model that makes reasonable predictions about which case marker to generate given the content words of a sentence, is expected to help MT and generation in general, particularly when the source (or native) and the target languages are morphologically divergent.

But how reliably can we predict case markers in Japanese using the information that exists only in the sentence? Consider the example in Figure 1. This sentence contains two case markers, *kara* 'from' and *ni*, the latter not corresponding to any word in English. If we were to predict the case markers in this sentence, there are multiple valid answers for each decision, many of which correspond to different semantic relations. For example, for the first case marker slot in Figure 1 filled by *kara*, *wa* (topic marker), *ni* 'in' or no case marker at all are all reasonable choices, while other markers such as *wo* (object marker), *de* 'at', *made* 'until', etc. are not considered reasonable. For the second slot filled by *ni*, *ga* (subject marker) is also a grammatically reasonable choice, making *Einstein* the subject of *idolize*, thus changing the meaning of the sentence. As is obvious in this example, the choice among the correct answers is determined by the speaker's intent in uttering the sentence, and is therefore impossible to recover from the content words or the sentence structure alone. At the same time, many impossible or unlikely case marking decisions can be eliminated by a case prediction model. Combined with an external component (for example an MT component) that can resolve semantic and intentional ambiguity, a case prediction model can be quite useful in sentence generation.

This paper discusses the task of case marker assignment in two distinct but related settings. After defining the task in Section 2 and describing our models in Section 3, we first discuss the *monolingual* task in Sections 4, whose goal is to predict the case markers

[1] Author names arranged alphabetically

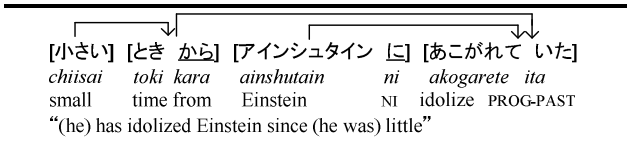

Figure 1. Example of case markers in Japanese (taken from the Kyoto Corpus). Square brackets indicate bunsetsu (phrase) boundaries, to be discussed below. Arrows between phrases indicate dependency relations.

using Japanese sentences and their dependency structure alone. We formulated this task after the well-studied task of semantic role labeling in English (e.g., Gildea and Jurafsky, 2002; Carreras and Màrques, 2005), whose goal is to assign one of 20 semantic role labels to each phrase in a sentence with respect to a given predicate, based on the annotations provided by PropBank (Palmer et al., 2005). Though the task of case marker prediction is more ambiguous and subject to uncertainty than the semantic role labeling task, we obtained some encouraging results which we present in Section 4. Next, in Section 5, we describe the *bilingual* task, in which information about case assignment can be extracted from a corresponding source language sentence. Though the process of MT introduces uncertainties in generating the features we use, we show that the benefit of using dependency structure in our models is far greater than not using it even when the assigned structure is not perfect.

2 The task of case prediction

In this section, we define the task of case prediction. We start with the description of the case markers we used in this study.

2.1 Nominal particles in Japanese

Traditionally, Japanese nominal postpositions are classified into the following three categories (e.g., Teramura, 1991; Masuoka and Takubo, 1992):

Case particles (or case markers). They indicate grammatical relations of the complement NP to the predicate. As they are jointly determined by the NP and the predicate, case markers often do not allow a simple mapping to a word in another language, which makes their generation more difficult. The relationship between the case marker and the grammatical relation it indicates is not straightforward either: a case marker can (and often does) indicate multiple grammatical relations as in *Ainshutain-ni akogareru* "idolize Einstein" where *ni* marks the Object relation, and in *Tokyo-ni sumu* "live in Tokyo" where *ni* indicates Location. Conversely, the same grammatical relation may be indicated by different case markers: both *ni* and *de* in *Tokyo-ni sumu* "live in Tokyo" and *Tokyo-de au* "meet in Tokyo" indicate the Location relation. We included 10 case markers as the primary target of prediction, as shown in the first 10 lines of Table 1.

Conjunctive particles. These particles are used to conjoin words and phrases, corresponding to English "and" and "or". As their occurrence is not predictable from the sentence structure alone, we did not include them in the current prediction task.

Focus particles. These particles add focus to a phrase against a given background or contextual knowledge, for example *shika* and *mo* in *pasuta-shika tabenakatta* "ate only pasta" and *pasuta-mo tabeta* "also ate pasta", corresponding to *only* and *also* respectively. Note that they often replace case markers: in the above examples, the object marker *wo* is no longer present when *shika* or *mo* is used. As they add information to the predicate-argument structure and are in principle not predictable given the sentence structure alone, we did not consider them as the target of our task. One exception is the topic marker *wa*, which we included as a target of prediction for the following reasons:

- Some linguists recognize *wa* as a topic marker, separately from other focus particles (e.g. Masuoka and Takubo, 1992). The main function of *wa* is to introduce a topic in the sentence, which is to a some extent predictable from the structure of the sentence.
- *wa* is extremely frequent in Japanese text. For example, it accounts for 13.2% of all postpositions in Kyoto University Text Corpus (henceforth Kyoto Corpus, Kurohashi and Nagao, 1997), making it the third most frequent postposition after *no* (20.57%) and *wo* (13.5%). Generating *wa* appropriately thus greatly enhances the readability of the text.
- Unlike other focus particles such as *shika* and *mo*, *wa* does not translate into any word in English, which makes it difficult to generate by using the information from the source language.

Therefore, in addition to the 10 true case markers, we also included *wa* as a case marker in our study.[2] Furthermore, we also included the combination of case particles plus *wa* as a secondary target of prediction. The case markers that can appear followed by *wa* are indicated by a check mark in the column "+*wa*" in Table 1. Thus there are seven secondary targets: *niwa, karawa, towa, dewa, ewa, madewa, yoriwa*. Therefore, we have in total 18 case particles to assign to phrases.

2.2 Task definition

The case prediction task we are solving is as follows. We are given a sentence as a list of *bunsetsu* together

[2] This set comprises the majority (92.5%) of the nominal particles, while conjunctive and focus particles account for only 7.5% of the nominal particles in Kyoto Corpus.

case markers		grammatical functions (e.g.)	+*wa*
が	*ga*	subject; object	
を	*wo*	object; path	
の[4]	*no*	genitive; subject	
に	*ni*	dative object, location	✓
から	*kara*	source	✓
と	*to*	quotative, reciprocal, *as*	✓
で	*de*	location, instrument, cause	✓
へ	*e*	goal, direction	✓
まで	*made*	goal (up to, until)	✓
より	*yori*	source, object of comparison	✓
は	*wa*	topic	

Table 1. Case markers included in this study

with a dependency structure. For our monolingual experiments, we used the dependency structure annotation in the Kyoto Corpus; for our bilingual experiments, we used automatically derived dependency structure (Quirk et al., 2005). Each *bunsetsu* (or simply *phrase* in this paper) is defined as consisting of one content word (or n-content words in the case of compounds with n-components) plus any number of function words (including particles, auxiliaries and affixes). Case markers are classified as function words, and there is at most one case marker per phrase.[3] In testing, the case marker for each phrase is hidden; the task is to assign to each phrase one of the 18 case markers defined above or NONE; NONE indicates that the phrase does not have a case marker.

2.3 Related work

Though the task of case marker prediction as formulated in this paper is novel, similar tasks have been defined in the past. The semantic role labeling task mentioned in Section 1 is one example; the task of function tag assignment in English (e.g., Blaheta and Charniak, 2000) is another. These tasks are similar to the case prediction task in that they try to assign semantic or function tags to a parsed structure. However, there is one major difference between these tasks and the current task: semantic role labels and function tags can for the most part be uniquely determined given the sentence and its parse structure; decisions about case markers, on the other hand, are highly ambiguous given the sentence structure alone, as mentioned in Section 1. This makes our task more ambiguous than the related tasks. As a concrete comparison, the two most frequent semantic role labels (ARG0 and ARG1) account for 60% of the labeled arguments in PropBank (Carreras and Màrquez, 2005), whereas our 2 most frequent case markers (*no* and *wo*) account for only 43% of the case-marked phrases. We should also note that semantic role labels and function tags have been artificially defined in accordance with theoretical decisions about what annotations should be useful for natural language understanding tasks; in contrast, the case markers are part of the surface sentence string and do not reflect any theoretical decisions.

The task of case prediction in Japanese has previously focused on recovering *implicit* case relations, which result when noun phrases are relativized or topicalized (e.g., Baldwin, 2000; Kawahara et al., 2004; Murata and Isahara, 2005). Their goal is different form ours, as we aim to generate surface forms of case markers rather than recover deeper case relations for which surface case marker are often used as a proxy.

In the context of sentence generation, Gamon et al. (2002) used a decision tree to classify nouns into one of the four cases in German, as part of their sentence realization from a semantic representation, achieving high accuracy (87% to 93.5%). Again, this is a substantially easier task than ours, because there are only four classes and one of them (nominative) accounts for 70% of all cases. Uchimoto et al. (2002), which is the work most related to ours, propose a model of generating function words (not limited to case markers) from "keywords" or headwords of phrases in Japanese. The components of their model are based on n-gram language models using the surface word strings and bunsetsu dependency information, and the results they report are not comparable to ours, as they limit their test sentences to the ones consisting only of two or three content words. We will see in the next section that our models are also quite different from theirs as we employ a much richer set of features.

3 Classifiers for case prediction

We implemented two types of models for the task of case prediction: *local models*, which choose the case marker of each phrase independently of the case markers of other phrases, and *joint models*, which incorporate dependencies among the case markers of dependents of the same head phrase. We describe the two types of models in turn.

3.1 Local classifiers

Following the standard practice in semantic role labeling, we divided the case prediction task into the tasks of *identification* and *classification* (Gildea and Jurafsky, 2002; Pradhan et al., 2004). In the identification task, we assign to each phrase one of two labels: HAS-CASE, meaning that the phrase has a case marker, or NONE, meaning that it does not have a case. In the

[3] One exception is that *no* can appear after certain case markers; in such cases, we considered *no* to be the case for the phrase.

[4] *no* is typically not considered as a case marker but rather as a conjunctive particle indicating adnominal relation; however, as *no* can also be used to indicate the subject in a relative clause, we included it in our study.

Basic features for phrases (self, parent)
HeadPOS, PrevHeadPOS, NextHeadPOS
PrevPOS,Prev2POS,NextPOS,Next2POS
HeadNounSubPos: time, formal nouns, adverbial
HeadLemma
HeadWord, PrevHeadWord, NextHeadWord
PrevWord, Prev2Word, NextWord, Next2Word
LastWordLemma (excluding case markers)
LastWordInfl (excluding case markers)
IsFiniteClause
IsDateExpression
IsNumberExpression
HasPredicateNominal
HasNominalizer
HasPunctuation: comma, period
HasFiniteClausalModifier
RelativePosition: sole, first, mid, last
NSiblings (number of siblings)
Position (absolute position among siblings)
Voice: pass, caus, passcaus
Negation
Basic features for phrase relations (parent-child pair)
DependencyType: D,P,A,I
Distance: linear distance in bunsetsu, 1, 2-5, >6
Subcat: POS tag of parent + POS tag of all children + indication for current
Combined features (selected)
HeadPOS + HeadLemma
ParentLemma + HeadLemma
Position + NSiblings
IsFiniteClause + GrandparentNounSubPos

Table 2: Basic and combined features for local classifiers

classification task, we assign one of the 18 case markers to each phrase that has been labeled with HASCASE by the identification model.

We train a binary classifier for identification and a multi-class classifier (with 18 classes) for classification. We obtain a classifier for the complete task by chaining the two classifiers. Let $P_{ID}(c|b)$ and $P_{CLS}(c|b)$ denote the probability of class c for bunsetsu b according to the identification and classification models, respectively. We define the probability distribution over classes of the complete model for case assignment as follows:

$$P_{CaseAssign}(NONE\ |b) = P_{ID}(NONE\ |b)$$

$$P_{CaseAssign}(l|b) = P_{ID}(HASCASE\ |b) * P_{CLS}(l|b)$$

Here, l denotes one of the 18 case markers.

We employ this decomposition mainly for efficiency in training: that is, the decomposition allows us to train the classification models on a subset of training examples consisting only of those phrases that have a case marker, following Toutanova et al. (2005). Among various machine learning methods that can be used to train the classifiers, we chose log-linear models for both identification and classification tasks, as they produce probability distributions which allows chaining of the two component models and easy integration into an MT system.

3.2 Joint classifiers

Toutanova et al. (2005) report a substantial improvement in performance on the semantic role labeling task by building a joint classifier, which takes the labels of other phrases into account when classifying a given phrase. This is motivated by the fact that the argument structure is a joint structure, with strong dependencies among arguments. Since the case markers also reflect the argument structure to some extent, we implemented a joint classifier for the case prediction task as well.

We applied the joint classifiers in the framework of *N*-best reranking (Collins, 2000), following Toutanova et al. (2005). That is, we produced *N*-best (*N*=5 in our experiments) case assignment sequence candidates for a set of sister phrases using the local models, and trained a joint classifier that learns to choose the best candidate from the set of sisters. The oracle accuracy of the 5-best candidate list was 95.9% per phrase.

4 Monolingual case prediction task

In this section we describe our models trained and evaluated using the gold-standard dependency annotations provided by the Kyoto Corpus. These annotations allow us to define a rich set of features exploring the syntactic structure.

4.1 Features

The basic local model features we used for the identification and classification models are listed in Table 2. They consist of features for a phrase, for its parent phrase and for their relations. Only one feature (GrandparentNounSubPos) currently refers to the grandparent of the phrase; all other features are between the phrase, its parent and its sibling nodes, and are a superset of the dependency-based features used by Hacioglu (2004) for the semantic labeling task. In addition to these basic features, we added 20 combined features, some of which are shown at the bottom of Table 2.

For the joint model, we implemented only two types of features: *sequence* of non-NONE case markers for a set of sister phrases, and *repetition* of non-NONE case markers. These features are intended to capture regularities in the sequence of case markers of phrases that modify the same head phrase.

All of these features are represented as binary features: that is, when the value of a feature is not binary, we have treated the combination of the feature name plus the value as a unique feature. With a count cut-off of 2 (i.e., features must occur at least twice to be in the model), we have 724,264 features in the identification

model, and 3,963,096 features in the classification model. The number of joint features in the joint model is 3,808. All models are trained using a Gaussian prior.

4.2 Data and baselines

We divided the Kyoto Corpus (version 3.0) into the following three sections:

- Training: contains news articles of January 1, 3-11 and editorial articles of January-August; 24,263 sentences, 234,474 phrases.
- Devtest: contains news articles of January 12-13 and editorial article of September. 4,833 sentences, 47,580 phrases.
- Test: contains news articles of January 14-17 and editorial articles of October-December. 9,287 sentences, 89,982 phrases.

The devtest set was used only for tuning model parameters and for performing error analysis.

As no previous work exists on the task of predicting case markers on the Kyoto Corpus, it is important to establish a good baseline. The simplest baseline of always selecting the most frequent label (NONE) gives us an accuracy of 47.5% on the test set. Out of the non-NONE case markers, the most frequent is *no*, which occurs in 26.6% of all case-marked phrases.

A more reasonable baseline is to use a language model to predict case. We trained and tested two language models: the first model, called KCLM, is trained on the same data as our log-linear models (24,263 sentences); the second model, called BigCLM, is trained on much more data from the same domain (826,373 sentences), taking advantage of the fact that language models do not require dependency annotation for training. The language models were trained using the CMU language modeling toolkit with default parameter settings (Clarkson and Rosenfeld, 1997).

We tested the language model baselines using the same task set-up as for our classifier: for each phrase, each of the 18 possible case markers and NONE is evaluated. The position for insertion of a case marker in each phrase is given according to our task set-up, i.e., at the end of a phrase preceding any punctuation. We choose the case assignment of the sequence of phrases in the sentence that maximizes the language model probability of the resulting sentence. We computed the most likely case assignment sequence using a dynamic programming algorithm.

4.3 Results and discussion

The results of running our models on case marker prediction are shown in Table 3. The first three rows correspond to the components of the local model: the identification task (*Id*, for all phrases), the classification task (*Cls*, only for case-marked phrases) and the complete task (*Both*, for all phrases). The accuracy on the complete task using the local model is 83.9%; the joint model improves it to 84.3%.

Models	Task	Training	Test
log-linear	Id	99.8	96.9
log-linear	Cls	96.6	74.3
log-linear (local)	Both	98.0	**83.9**
log-linear(joint)	Both	97.8	**84.3**
baseline (frequency)	Both	48.2	47.5
baseline (KCLM)	Both	93.9	67.0
baseline (BigCLM)	Both	—	78.0

Table 3: Accuracy of case prediction models (%)

The improvement due to the joint model is small in absolute percentage points (0.4%), but is statistically significant according to a test for the difference of proportions ($p< 0.05$). The use of a joint classifier did not lead to as large an improvement over the local classifier as for the semantic role labeling task. There are several reasons for that we can think of. First, we have only used a limited set of features for the joint model, i.e., case sequence and repetition features. A more extensive use of global features might lead to a larger improvement. Secondly, unlike the task of semantic role labeling, where there are about 20 phrases that need to be labeled with respect to a predicate, about 50% of all phrases in the Kyoto Corpus do not have sister nodes. This means that these phrases cannot take advantage of the joint classifier using the current model formulation. Finally, case markers are much shallower than semantic role labels in the level of linguistic analysis, and so are inherently subject to more variations, including missing arguments (so called zero pronouns) and repeated case markers corresponding to different semantic roles.

From Table 3, it is clear that our models outperform the baseline model significantly. The language model trained on the same data has much lower performance (67.0% vs. 84.3%), which shows that our system is exploiting the training data much more efficiently by looking at the dependency and other syntactic features. An inspection of the 500 most highly weighted features also indicates that phrase dependency-based features are very useful for both identification and classification. Given much more data, though, the language model improves significantly to 78%, but our classifier still achieves a 29% error reduction over it. The differences between the language models and the log-linear models are statistically significant at level $p < 0.01$ according to a test for the difference of proportions.

Figure 2 plots the recall and precision for the frequently occurring (>500) cases. We achieve good results on NONE and *no*, which are the least ambiguous decisions. Cases such as *ni*, *wa*, *ga*, and *de* are highly confusable with other markers as they indicate multiple grammatical relations, and the performance of our

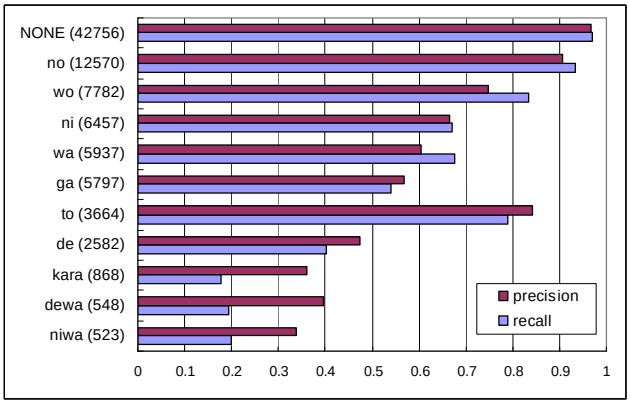

Figure 2: Precision and recall per case marker (frequency in parentheses)

models on them is therefore limited. As expected, performance (especially recall) on secondary targets (*dewa*, *niwa*) suffers greatly due to the ambiguity with their primary targets.

5 Bilingual case prediction task: simulating case prediction in MT

Incorporating a case prediction model into MT requires taking additional factors into consideration, compared to the monolingual task described above. On the one hand, we need to extend our model to handle the additional knowledge source, i.e., the source sentence. This can potentially provide very useful features to our model, which are not available in the monolingual task. On the other hand, since gold-standard dependency annotation is not available in the MT context, we must deal with the imperfections in structural annotations.

In this section, we describe our case prediction models in the context of English-to-Japanese MT. In this setting, dependency information for the target language (Japanese) is available only through projection of a dependency structure from the source language (English) in a tree-to-string-based statistical MT system (Quirk et al., 2005). We conducted experiments using the English source sentences and the reference translations in Japanese: that is, our task is to predict the case markers of the Japanese reference translations correctly using all other words in the reference sentence, information from the source sentence through word alignment, and the Japanese dependency structure projected via an MT component. Ultimately, our goal is to improve the case marker assignment of a candidate translation using a case prediction model; the experiments described in this section on reference translations serve as an important preliminary step toward achieving that final goal. We will show in this section that even the automatically derived syntactic information is very useful in assigning case markers in the target language, and that utilizing the information from the source language also greatly contributes to reducing case marking errors.

5.1 Data and task set-up

The dataset we used is a collection of parallel English-Japanese sentences from a technical (computer) domain. We used 15,000 sentence pairs for training, 5,000 for development, and 4,241 for testing.

The parallel sentences were word-aligned using GIZA++ (Och and Ney, 2000), and submitted to a tree-to-string-based MT system (Quirk et al., 2005) which utilizes the dependency structure of the source language and projects dependency structure to the target language. Figure 3 shows an example of an aligned sentence pair: on the source (English) side, part-of-speech (POS) tags and word dependency structure are assigned (solid arcs). The alignments between English and Japanese words are indicated by the dotted lines. In order to create phrase-level dependency structures like the ones utilized in the Kyoto Corpus monolingual task, we derived some additional information for the Japanese sentence in the following manner.

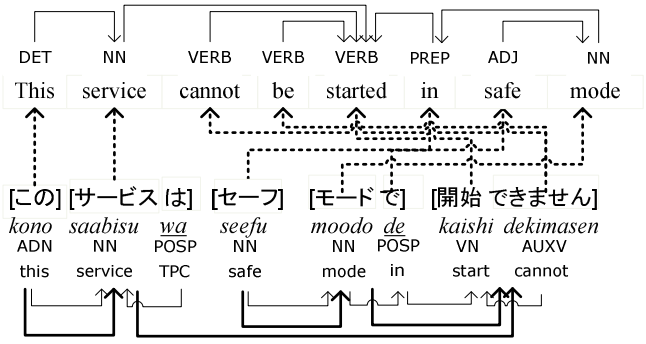

Figure 3. Aligned English-Japanese sentence pair

First, we tagged the sentence using an automatic tagger with a set of 19 POS tags. We used these POS tags to parse the words into phrases (bunsetsu): each bunsetsu consists of one content word plus any number of function words, where content and function words are defined via POS. We then constructed a phrase-level dependency structure using a breadth-first traversal of the word dependency structure projected from English. These phrase dependencies are indicated by bold arcs in Figure 3. The case markers to be predicted (*wa* and *de* in this case) are underlined.

The task of case marker prediction is the same as described in Section 2: to assign one of the 18 case markers described in Section 2 or NONE to each phrase.

5.2 Baseline models

We implemented the baseline models discussed in Section 4.2 for this domain as well. The most frequent

Monolingual features	
Feature	Example
HeadWord /HeadPOS	*saabisu/NN*
PrevWord/PrevPOS	*kono/AND*
Prev2Word/Prev2WordPOS	*none/none*
NextWord/NextPOS	*seefu/NN*
Next2Word/Net2POS	*moodo/NN*
PrevHeadWord/PrevHeadPOS	*kono/AND*
NextHeadWord/NextHeadPOS	*seefu/NN*
ParentHeadWord/ParentHeadPOS	*kaishi/VN*
Subcat: POS tags of all sisters and parent	*NN-c,NN,VN-h*
NSiblings (including self)	2
Distance	1
Direction	*left*
Alternative Parent Word /POS	*saabisu/NN*
Bilingual features	
Feature	Example
Word/POS of source words aligned to the head of the phrase	*service/NN*
Word/POS of all source words aligned to any word in the phrase	*service/NN*
Word/POS of all source words aligned to the head word of the parent phrase	*started/VERB*
Word/POS of all source words aligned to alternative parent words of the phrase	*service/NN, started/VERB*
All source preposition words	*in*
Word/POS of parent of source word aligned to any word in the phrase	*started/VERB*
Aligned Subcat	*NN-c,VERB,VERB,VERB-h,PREP*
Aligned NSiblings	4
Aligned Distance	2
Aligned Direction	*left*

Table 4: Monolingual and bilingual features

case assignment is again NONE, which accounts for 62.0% of the test set. The frequency of NONE is higher in this task than in the Kyoto Corpus, because our bunsetsu-parsing algorithm prefers to err on the side of making too many rather than too few phrases. This is because our final goal is to generate all case markers, and if we mistakenly joined two bunsetsu into one, our case assigner would be able to propose only one case marker for the resulting bunsetsu, which would be necessarily wrong if both bunsetsu had case markers. The most frequent case marker is again *no*, which occurs in 29.4% of all case-marked phrases. As in the monolingual task, we trained two trigram language models: one was trained on the training set of our case prediction models (15,000 sentences); another was trained on a much larger set of 450,000 sentences from the same domain. The results of these baselines are discussed in Section 5.4.

5.3 Log-linear models

The models we built for this task are log-linear models as described in Section 3. In order to isolate the impact of information from the source language available for the case prediction task, we built two kinds of models:

Model	Test data
baseline (frequency)	62.0
baseline (15kLM)	79.0
baseline (450kLM)	83.6
log-linear monolingual	**85.3**
log-linear bilingual	**92.3**

Table 5: Accuracy of bilingual case prediction (%)

monolingual models, which do not use any information from the source English sentences, and *bilingual* models, which use information from the source. Both models are local models in the sense discussed in Section 3.

Table 4 shows the features used in the monolingual and bilingual models, along with the examples (the value of the feature for the phrase [*saabisu wa*] in Figure 3); in addition to these, we also provided some feature combinations for both monolingual and bilingual models. Many of the monolingual features (i.e., first 11 lines in Table 4) are also present in Table 2. Note that lexically based features are of greater importance for this task, as the dependency information available in this context is of much poorer quality than that provided by the Kyoto Corpus. In addition to the features in Table 2, we added a Direction feature (with values *left* and *right*), and an Alternative Parent feature. Alternative parents are all words which are the parents of any word in the phrase, according to the word-based dependency tree, with the constraint that case markers cannot be alternative parents. This feature captures the information that is potentially lost in the process of building a phrase dependency structure from word dependency information in the target language.

The bottom half of Table 4 shows bilingual features. The features of the source sentence are obtained through word alignments. We create features from the source words aligned to the head of the phrase, to the head of the parent phrase, or to any alterative parents. If any word in the phrase is aligned to a preposition in the source language, our model can use the information as well. In addition to word- and POS-features for aligned source words, we also refer to the corresponding dependency between the phrase and its parent phrase in the English source. If the head of the Japanese phrase is aligned to a single source word s_1, and the head of its parent phrase is aligned to a single source word s_2, we extract the relationship between s_1 and s_2, and define subcategorization, direction, distance, and number of siblings features, in order to capture the grammatical relation in the source, which is more reliable than in the projected target dependency structure.

5.4 Results and discussion

Table 5 summarizes the results on the complete case assignment task in the MT context. Compared to the language model trained on the same data (15kLM), our

monolingual model performs significantly better, achieving a 30% error reduction (85.3% vs. 79.0%). Our monolingual model outperforms even the language model trained on 30 times more data (85.3% vs. 83.6%), with an error reduction of 10%. The difference is statistically significant at level $p < 0.01$ according to a test for the difference of proportions. This means that even though the projected dependency information is not perfect, it is still useful for the case prediction task.

When we add the bilingual features, the error rate of our model is cut almost in half: the bilingual model achieves an error reduction of 48% over the monolingual model (92.3% vs. 85.3%, statistically significant at level $p < 0.01$). This result is very encouraging: it indicates that information from the source sentence can be exploited very effectively to improve the accuracy of case assignment. The usefulness of the source language information is also obvious when we inspect which case markers had the largest gains in accuracy due to this information: the top three cases were *kara* (0.28 to 0.65, a 57% gain), *dewa* (0.44 to 0.65, a 32% gain) and *to* (0.64 to 0.85, a 24% gain), all of which have translations as English prepositions. Markers such as *ga* (subject marker, 0.68 to 0.74, a 8% gain) and *wo* (object marker, 0.83 to 0.86, a 3.5% gain), on the other hand, showed only a limited gain.

6 Conclusion and future directions

This paper described the task of predicting case markers in Japanese, and reported results in a monolingual and a bilingual settings. The results show that the models we proposed, which explore syntax-based features and features from the source language in the bilingual task, can effectively predict case markers.

There are a number of extensions and next steps we can think of at this point, the most immediate and important one of which is to incorporate the proposed model in an end-to-end MT system to make improvements in the output of MT. We would also like to perform a more extensive analysis of features and feature ablation experiments. Finally, we would also like to extend the proposed model to include languages with inflectional morphology and the prediction of grammatical elements in general.

Acknowledgements

We would like to thank the anonymous reviewers for their comments, and Bob Moore, Arul Menezes, Chris Quirk, and Lucy Vanderwende for helpful discussions.

References

Baldwin, T. 2004. Making Sense of Japanese Relative Clause Constructions, In *Proceedings of the 2nd Workshop on Text Meaning and Interpretation*.

Blaheta, D. and E. Charniak. 2000. Assigning function tags to parsed text. In *Proceedings of NAACL*, pp.234-240.

Carreras, X. and L. Màrquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of CoNLL-2005*.

Clarkson, P.R. and R. Rosenfeld. 1997. Statistical Language Modeling Using the CMU-Cambridge Toolkit. In *Proceedings of ESCA Eurospeech*, pp. 2007-2010.

Collins, M. 2000. Discriminative reranking for natural language parsing. In *Proceedings of ICML*.

Gamon, M., E. Ringger, S. Corston-Oliver and R. Moore. 2002. Machine-learned Context for Linguistic Operations in German Sentence Realization. In *Proceeding of ACL*.

Gildea, D. and D. Jurafsky. 2002. Automatic Labeling of Semantic Roles. In *Computational Linguistics* 28(3): 245-288.

Hacioglu, K. 2004. Semantic Role Labeling using Dependency Trees. In *Proceedings of COLING 2004*.

Kawahara, D., N. Kaji and S. Kurohashi. 2000. Japanese Case Structure Analysis by Unsupervised Construction of a Case Frame Dictionary. In *Proceedings of COLING*, pp. 432-438.

Kurohashi, S. and M.Nagao. 1997. Kyoto University Text Corpus Project. In *Proceedings of ANLP*, pp.115-118.

Masuoka, T. and Y. Takubo. 1992. *Kiso Nihongo Bunpou* (Fundamental Japanese grammar), revised version. Kuroshio Shuppan, Tokyo.

Murata, M., and H. Isahara. 2005. Japanese Case Analysis Based on Machine Learning Method that Uses Borrowed Supervised Data. In Proceedings of *IEEE NLP-KE-2005*, pp.774-779.

Och, F.J. and H. Ney. 2000. Improved statistical alignment models. In *Proceedings of ACL*: pp.440-447.

Palmer, M., D. Gildea and P. Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. In *Computational Linguistics* 31(1).

Pradhan, S., W. Ward, K. Hacioglu, L. Martin, D. Jurafsky. 2004. Shallow Semantic Parsing Using Support Vector Machines. In *Proceedings of HLT/NAACL*.

Quirk, C., A. Menezes and C. Cherry. 2005. Dependency Tree Translation: Syntactically Informed Phrasal SMT. In *Proceedings of ACL*.

Teramura, H. 1991. *Nihongo-no shintakusu-to imi* (Japanese syntax and meaning). Volume III. Kuroshio Shuppan, Tokyo.

Toutanova, K., A. Haghighi and C. D. Manning. 2005. Joint Learning Improves Semantic Role Labeling. In *Proceeding of ACL*, pp.589-596.

Uchimoto, K., S. Sekine and H. Isahara. 2002. Text Generation from Keywords. In *Proceedings of COLING 2002*, pp.1037-1043.