A Discriminative Approach to Japanese Zero Anaphora Resolution with Large-scale Lexicalized Case Frames

Ryohei Sasano
Precision and Intelligence Laboratory
Tokyo Institute of Technology
sasano@pi.titech.ac.jp

Sadao Kurohashi
Graduate School of Informatics
Kyoto University
kuro@i.kyoto-u.ac.jp

Abstract

We present a discriminative model for Japanese zero anaphora resolution that simultaneously determines an appropriate case frame for a given predicate and its predicate-argument structure. Our model is based on a log linear framework, and exploits lexical features obtained from a large raw corpus, as well as non-lexical features obtained from a relatively small annotated corpus. We report the results of zero anaphora resolution on Web text and demonstrate the effectiveness of our approach. In addition, we also investigate the relative importance of each feature for resolving zero anaphora in Web text.

1 Introduction

Zero anaphora resolution is the task of detecting and identifying the omitted arguments of a predicate. Since arguments are often omitted in Japanese, zero anaphora resolution plays an important role in the analysis of Japanese predicate-argument structures. For example, in the following text:

(i) *Musuko-wa*	*itazura-ga*	*sukide*	
son-TOP	mischief-NOM	like	
watashi-mo	*(ϕ-ni)*	*te-wo*	*yaiteiru.*
I	ϕ-DAT	hands-ACC	burn (have difficulty)

(My son likes mischief, so I have difficulty with ϕ.)

the dative argument of the predicate '*yaku*[1] (burn)' has been omitted. The omitted element is called a zero pronoun, and in this example it refers to '*musuko* (son).' Although most previous work has focused on zero anaphora in newspaper articles, this paper aims to resolve zero anaphora in Web text, since this involves a wider range of writing styles and is thus considered to be a more practical setting.

Reference resolution systems generally require a variety of information sources ranging from syntactic and discourse preferences to semantic preferences (Ng and Cardie, 2002; Haghighi and Klein, 2010). Since syntactic and discourse preferences are not word-specific, they can be learned from a relatively small annotated corpus. Semantic preferences, on the other hand, represent highly lexicalized knowledge, and hence it is difficult to learn these from a small annotated corpus. In some cases, knowledge of the relations between a predicate and its particular argument is insufficient, particularly for zero anaphora resolution. For example, although the dative argument of the predicate '*yaku* (bake/burn)' is generally filled by a *disk*, such as a CD or DVD, it is often filled by a *person*, such as '*musuko* (son),' if the accusative argument is filled by '*te* (hands)' as in the example (i).[2] Thus, knowledge of relations among a predicate and its multiple arguments is required to take such preferences into account.

Sasano et al. (2008) exploited large-scale case frames that were automatically constructed from 1.6 billion Web sentences as such a lexical resource, and proposed a probabilistic model for Japanese zero anaphora resolution. Their model demonstrated moderate performance, but it could not easily introduce new features, especially overlapping ones, nor take into consideration the importance of each feature, due to the assumption of independence in estimating probability. However, we think a variety of clues can be useful for zero anaphora resolution, where it is important to exploit overlapping features and consider the importance of each feature.

[1] '*Yaku*' is the original form of '*yaiteiru*.'

[2] '*Te-wo yaku*' (literally 'burn hands') is a Japanese idiom, which means 'have difficulty' in English.

Therefore, in this paper, we extend Sasano et al. (2008)'s model by incorporating it into a log-linear framework, and introduce overlapping features such as lexical features with different granularities. In addition, we also investigate the relative importance of each feature for resolving zero anaphora in Web text.

2 Related Work

Several approaches to Japanese zero anaphora resolution have been proposed. Seki et al. (2002) proposed a probabilistic model for zero pronoun detection and resolution that used hand-crafted case frames. Kawahara and Kurohashi (2004) introduced wide-coverage case frames that were automatically constructed from a large corpus to alleviate the sparseness of hand-crafted case frames. They used the case frames as selectional restrictions for zero pronoun resolution. Iida et al. (2006) explored a machine learning method using rich syntactic pattern features that represented the syntactic relations between a zero-pronoun and its candidate antecedent.

Since predicate-argument structure analysis and zero anaphora resolution are closely related, several approaches have simultaneously solved these two tasks. Sasano et al. (2008) proposed a lexicalized probabilistic model for zero anaphora resolution, which adopted an entity-mention model and simultaneously resolved predicate-argument structures and zero anaphora. Taira et al. (2008) proposed a model for analyzing predicate-argument structures by using decision lists, which integrated the tasks of semantic role labeling and zero-pronoun identification. Imamura et al. (2009) proposed a discriminative model for analyzing predicate-argument structures that simultaneously conducted zero anaphora resolution.

For languages other than Japanese, Ferrandez and Peral (2000) proposed a hand-engineered rule-based method for both determining anaphoricity and identifying antecedents in Spanish zero pronoun resolution. Zhao and Ng (2007) proposed feature-based methods to Chinese zero anaphora resolution. Kong and Zhou (2010) proposed a tree kernel-based unified framework for Chinese zero anaphora resolution, which dealt with three subtasks: zero anaphora detection, anaphoricity determination, and antecedent identification.

3 Case Frames

3.1 Lexicalized case frames

Our model exploits lexicalized case frames that are automatically constructed from 1.6 billion Web sentences by using Kawahara and Kurohashi (2002)'s method. Case frames are constructed for each predicate like PropBank frames (Palmer et al., 2005), and for each meaning of the predicate like FrameNet frames (Fillmore et al., 2003). However, neither pseudo-semantic role labels such as Arg1 in PropBank nor information about frames defined in FrameNet are included in the case frames. Each case frame describes surface cases that each predicate has and examples that can fill a case slot, which is fully-lexicalized like the subcategorization lexicon VALEX (Korhonen et al., 2006).

Note that case frames offer not only knowledge of the relations between a predicate and its particular case slot, but also knowledge of the relations among a predicate and its multiple case slots. Table 1 shows examples of constructed case frames.[3] A different case frame is constructed for each meaning of '*yaku* (bake/burn),' such as 'bake foods,' 'have difficulty,' and 'burn on a disk.'

3.2 Generalization of examples

The data sparseness problem is alleviated to some extent but not eliminated by using case frames that are automatically constructed from a large corpus. For instance, there are thousands of named entities (NEs) that cannot intrinsically be covered. Sasano et al. (2008) generalized examples of case slots based on 22 common noun categories defined in the Japanese morphological analyzer JUMAN,[4] and 8 NE classes defined by the IREX Committee (1999) to deal with this problem.

In addition, we generalized case slot examples based on automatically acquired multi-word noun clusters. Kazama and Torisawa (2008) proposed the parallelization of EM-based clustering with the aim of enabling large-scale clustering and using the resulting clusters in NE recognition. We used the resulting 2,000 clusters acquired from 1 million unique multi-word nouns.

Table 2 lists examples of the resulting 2,000 clusters.[3] As well as common noun categories and NE classes, we calculated the average of the probabilities that each case slot example belonged to

[3] We use English examples for the sake of readability.

[4] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html

	Case slot	Examples:#	Generalized examples:%
yaku (1) (bake)	*ga* nominative	I:39, owner:26, daughter:22, mother:19, $\cdots$, Asako:2, $\cdots$	[CT:PERSON]:0.620, [NE:PERSON]:0.116, [CL:887]:0.070, $\cdots$
	wo accusative	bakery:9265, cake:4495, meat:4057, fish:2002, $\cdots$	[CT:FOOD]:0.711, [CL:883]:0.221, [CT:BODY PART]:0.105, $\cdots$
	ni dative	snack:35, breakfast:33, birthday:29, gift:21, $\cdots$	[CT:ABSTRACTION]:0.233, [CT:FOOD]:0.171, [CL:624]:0.076, $\cdots$
	de tools/methods	frying pan:894, moderate heat:425, $\cdots$	[CT:ARTIFACT]:0.356, [CL:291]:0.252, $\cdots$
		⋮	
yaku (2) (have difficulty)	*ga* nominative	who:7, teacher:7, everyone:5, family:4, government:3, $\cdots$	[CT:PERSON]:0.372, [NE:PERSON]:0.128, [CT:ORGANIZATION]:0.128, $\cdots$
	wo accusative	hand:6864	[CT:BODY PART]:1.000
	ni dative	child:52, attack:43, treatment:40, provision:32, daughter:30, $\cdots$	[CT:ABSTRACTION]:0.432, [CT:PERSON]:0.172, [NE:PERSON]:0.060, [CL:32]:0.016, $\cdots$
		⋮	
yaku (3) (burn)	*ga* nominative	I:1, husband:1,	[CT:PERSON]:1.000
	wo accusative	file:20, tune:14, music:9, image:8, video:4, $\cdots$	[CT:ABSTRACTION]:0.645, [CT:ARTIFACT]:0.273, $\cdots$
	ni dative	CD:3106, DVD:2066, $\cdots$	[CL:70]:0.829, $\cdots$
	de tools	machine:21, writing soft:10, PC:5, iTunes:4, $\cdots$	[CT:ABSTRACTION]:0.294, [CT:ARTIFACT]:0.191, [NE:ARTIFACT]:0.054, $\cdots$
		⋮	
		⋮	

Table 1: Example case frames for '*yaku* (bake / have difficulty / burn).'

Cluster	Nouns
CL:32	child (0.974), infant (0.738), kid (0.727), babies and infant (0.436), $\cdots$
CL:70	CD (0.896), DVD (0.837), CD-ROM (0.603), cassette (0.512), $\cdots$
CL:291	low heat (0.720), slow fire (0.715), moderate heat (0.681), distant fire (0.678), $\cdots$
CL:624	dinner (0.926), supper (0.925), lunch (0.882), breakfast (0.868), $\cdots$
CL:883	Chinese noodles (0.860), noodles (0.801), curry (0.793), cake (0.749), $\cdots$
CL:887	mother (0.909), parents (0.875), mom (0.838), husband (0.775), father (0.774), $\cdots$

Table 2: Examples of resulting 2,000 clusters (Kazama and Torisawa, 2008). Nouns that have high probabilities of belonging to target clusters are shown with probabilities.

the target cluster, and added it to the case slot. For example, if examples of a case slot include 'CD:3106,' 'DVD:2066,' and 271 other examples that have no probability of belonging to cluster 70, the average for cluster 70 is calculated as:

$$\frac{0.896 \times 3106 + 0.837 \times 2066}{3106 + 2066 + 271} \approx 0.829.$$

This type of generalized example represents more fine-grained semantic categories compared with examples that are generalized by using common noun categories and NE classes. The generalized examples are also included in Table 1, such as "[CL:70]:0.829" in the dative case slot of '*yaku* (3).' CT, NE, and CL in generalized examples denote common noun **ca**te**g**ory, **n**amed **e**ntity class, and multi-word noun **cl**uster, respectively.

4 A Discriminative Model for Zero Anaphora Resolution

4.1 Overview

Our model basically follows that of Sasano et al. (2008), except for the method of estimating possible combinations of case frames and predicate-argument structures. We also limited the target cases for zero anaphora resolution to '*ga*' (nominative), '*wo*' (accusative), and '*ni*' (dative) cases. The outline of our model is as follows:

1. Parse an input text and recognize NEs.
2. Resolve coreference and link each mention to a discourse entity or create a new entity.
3. From the end of each sentence, analyze the predicate-argument structure for each verb or adjective using the following steps:

1.	$<cf=$‘*yaku*’(1), $s=$[NOM:‘*watashi*’ (I), ACC:‘*te*’ (hands), DAT:‘*musuko*’ (son)]$>$
2.	$<cf=$‘*yaku*’(1), $s=$[NOM:‘*watashi*’ (I), ACC:‘*te*’ (hands), DAT:`NULL`’]$>$
3.	$<cf=$‘*yaku*’(1), $s=$[NOM:`NULL`, ACC:‘*te*’ (hands), DAT:‘*watashi*’ (I)]$>$
4.	$<cf=$‘*yaku*’(1), $s=$[NOM:‘*musuko*’ (son), ACC:‘*te*’ (hands), DAT:`NULL`$>$
5.	$\boldsymbol{<cf=}$**‘*yaku*’(2),** $\boldsymbol{s=}$**[NOM:‘*watashi*’ (I), ACC:‘*te*’ (hands), DAT:‘*musuko*’ (son)]**$\boldsymbol{>}$
6.	$<cf=$‘*yaku*’(2), $s=$[NOM:‘*watashi*’ (I), ACC:‘*te*’ (hands), DAT:`NULL`’]$>$
	⋮

Table 3: Examples of possible combinations of case frame cf and predicate-argument structure s for the predicate ‘*yaku*’ in the example (i) in Section 1. Bold font indicates the proper combination for this example.

(a) Select a case frame temporarily.

(b) List possible predicate-argument structures including omitted arguments.

(c) Estimate possible combinations of case frames and predicate-argument structures, and select the one with the highest estimate.

In 3-(b), we first consider only the overt arguments and prune away improbable structures to reduce the search space. We apply a log-linear framework to estimating a combination of a case frame and a predicate-argument structure to introduce overlapping features and take into consideration the relative importance of each feature.

Note that the estimation is not separately conducted for each argument, but for all arguments including overt and omitted arguments. For examples, when we analyze the predicate ‘*yaku*’ in the example (i) in Section 1, we consider the various combinations as listed in Table 3, and choose the combination with the highest estimate.

4.2 Log-linear model

When text t and target predicate p are given, we choose the combination of case frame cf and predicate-argument structure s with the highest conditional probability,

$$(cf_{best}, s_{best}) = \arg\max_{cf,s} P(cf, s|p, t).$$

We model the conditional probability, using a log-linear framework:

$$P(cf, s|p, t; \Lambda) = \frac{1}{Z(p,t)} \exp\{\Lambda \cdot \boldsymbol{F}(cf, s, p, t)\},$$

$$Z(p,t) = \sum_{\{cf,s\} \in C(p,t)} \exp\{\Lambda \cdot \boldsymbol{F}(cf, s, p, t)\},$$

where $\boldsymbol{F} = (f_1, \ldots, f_K)$ is a feature vector whose elements represent K feature functions, $\Lambda = (\lambda_1, \ldots, \lambda_K)$ denotes a weight vector (parameter vector) for the feature functions, and $C(p,t)$ yields a set of possible combinations of case frame cf and predicate-argument structure s for given predicate p and text t.

4.3 Parameter estimation

We now describe how parameter vector Λ is estimated from a set of training data. When the training set consisting of N instances $\{(s^{(1)}, p^{(1)}, t^{(1)}), (s^{(2)}, p^{(2)}, t^{(2)}), \ldots, (s^{(N)}, p^{(N)}, t^{(N)})\}$ is given, we choose the combination of CF and Λ that maximize the posterior probability:

$$\max_{CF,\Lambda} \left\{ \sum_{n=1}^{N} \log P(cf^{(n)}, s^{(n)}|p^{(n)}, t^{(n)}; \Lambda) - \alpha||\Lambda||^2 \right\},$$

where α is a regularization parameter for the L2 norm, and $CF = (cf^{(1)}, cf^{(2)}, \ldots, cf^{(N)})$ is a combination of possible case frames, i.e. $cf^{(n)}$ is a candidate case frame for the given instance $(s^{(n)}, p^{(n)}, t^{(n)})$. Since the appropriate case frames are not annotated in the training set, we choose an appropriate case frame in estimating parameters with the following algorithm:

1. Initialize parameter Λ to a random value in the range [0,1].

2. For each training instance, update $cf^{(n)}$ that maximizes $P(cf, s|p, t; \Lambda)$ with current parameter Λ:

   $$\hat{cf}^{(n)} = \arg\max_{cf^{(n)}} P(cf^{(n)}, s^{(n)}|p^{(n)}, t^{(n)}; \Lambda)$$

   If $CF = (cf^{(1)}, cf^{(2)}, \ldots, cf^{(N)})$ is not updated, we determine current parameter Λ as the resulting parameter.

3. If CF is updated, we renew parameter Λ that maximizes the posterior probability with N training instances $\{(cf^{(1)}, s^{(1)}, p^{(1)}, t^{(1)}), \ldots, (cf^{(N)}, s^{(N)}, p^{(N)}, t^{(N)})\}$,

$$\mathcal{L}_\Lambda = \sum_{n=1}^{N} \log P(cf^{(n)}, s^{(n)} | p^{(n)}, t^{(n)}; \Lambda) - \alpha ||\Lambda||^2,$$

and go back to step 2. To avoid overfitting, we include an L2-regularization term in the objective.[5] $\mathcal{L}_\Lambda$ is maximized by the Limited memory BFGS (L-BFGS) algorithm (Nocedal, 1980).[6]

In both steps 2 and 3, the log-likelihood increases monotonically, and this algorithm thus always converges to an optimal solution but does not ensure the global maximum parameter will be assigned. That is, we can obtain convergence to different local optima. Therefore, we test several initial values and adopt the resulting parameter that maximizes posterior probability.

5 Features

5.1 Lexical features

We exploit six types of lexical features: word PMI (Pointwise Mutual Information), cluster PMI, category PMI, NE PMI, occupancy of a case slot, and overt argument assignment score. Their values are real values that are calculated by using case frames. Since our model is based on the entity-mention model that assigns zero pronouns not to a certain mention but to an entity, several values can be calculated for a certain lexical feature by taking coreferential mentions into consideration. In such cases, we choose the highest value as a corresponding lexical feature.

Word PMI Each case slot of a case frame has typical words that are often assigned to the slot. We use the PMI features between a slot of a case frame and its antecedent candidate to reflect such preferences:

e.g.

- $\log\{P(\textit{child}|\text{cf=}\textit{yaku}(2)\text{,case=}\textit{ni})/P(\textit{child})\}$

As well as most other features, this type of features is distinguished by the case of zero pronouns: '*ga*,' '*wo*,' and '*ni*,' respectively.

Cluster, Category, and NE PMI We also use generalized example versions of word PMI to alleviate the data sparseness problem in word PMI:

e.g.

- $\log\{P([\text{CL:32}]\ |\textit{yaku}(2), \textit{ni})/P([\text{CL:32}])\}$
- $\log\{P([\text{CT:PERSON}]|\textit{yaku}(2), \textit{ni})/P([\text{CT:PERSON}])\}$
- $\log\{P([\text{NE:PERSON}]|\textit{yaku}(2), \textit{ni})/P([\text{NE:PERSON}])\}$

After this, we will generically call these three features **GE PMI**. All the features described above overlap, and only their granularities are different.

Occupancy of case slot We believe that there is a relation between the occupancy of a case slot and its generativity of zero pronouns. For example, since the *ni* (dative) case of '*yaku* (3)' often appears, we assume this slot is just omitted as a zero pronoun even if there is no overt argument. However, since the *ni* (dative) case of '*yaku* (1)' rarely appears, we assume there is no case slot if there is no overt argument.

Therefore, we use the log occupancy of the case slot, whose value is the same as the log of the generative probability of a case slot in Kawahara and Kurohashi (2006)'s model and estimated from case structure analysis of a large raw corpus:

e.g.

- $\log P(A(cs) = 1|\text{cf=}\textit{yaku}(2)\text{,case=}\textit{ni})$,

  where $A(cs) = 1$ denotes that the target case slot is occupied by an overt argument.

Overt argument assignment score Our model not only takes into account the correspondence between a case slot and its omitted argument, but also the correspondence between a case slot and its overt argument. The score for overt argument assignment reflects the likelihood of an overt argument assignment, and this is the same as the score for the predicate-argument structure in Kawahara and Kurohashi (2006)'s model, which does not take into consideration zero anaphoric relations.

e.g.

- $\log P(\text{cf=}\textit{yaku}(2), \textit{ga}_{ov}\text{:}\textit{watashi}, \textit{wo}_{ov}\text{:}\textit{te}|\textit{yaku})$

Only this feature is not separately calculated for each case, but for the whole overt predicate-argument structure. Note that, this feature also includes non-lexical preferences in the analysis of a predicate-argument structure.

5.2 Non-lexical features

We also exploit three types of non-lexical features, which are binary features to reflect syntactic and

[5]We set α to 1.0 in all our experiments.

[6]We used libLBFGS 1.9: http://www.chokkan.org/software/liblbfgs/.

Intra sentential (64 categories)	
Itopic:	Mentioned with **topic** marker
IP-self:	Mentioned at **p**arent node
IC-self:	Mentioned at **c**hild node
IGP-self:	Mentioned at **g**rand-**p**arent node
IGC-self:	Mentioned at **g**rand-**c**hild node
⋮	⋮
IB-self:	Mentioned at a preceding node in the sentence except above (**b**efore)
IA-self:	Mentioned at a following node in the sentence except above (**a**fter)
IP-*ga*-ov:	Overt nominative argument of a predicate in **p**arent node
IP-*ga*-om:	Omitted nominative argument of a predicate in **p**arent node
IP-*wo*-ov:	Overt accusative argument of a predicate in **p**arent node
IP-*wo*-om:	Omitted accusative argument of a predicate in **p**arent node
⋮	⋮
IGP-*ga*-ov:	Overt nominative argument of a predicate in **g**rand-**p**arent node
⋮	⋮
Inter sentential (21 categories)	
B1:	Mentioned in the adjacent sentence
B1-*ga*-ov:	Overt nominative argument of a predicate in the adjacent sentence
B1-*ga*-om:	Omitted nominative argument of a predicate in the adjacent sentence
B1-*wo*-ov:	Overt accusative argument of a predicate in the adjacent sentence
⋮	⋮
B2:	Mentioned in two sentences before
B2-*ga*-ov:	Overt nominative argument of a predicate in the two sentence before
⋮	⋮
B3:	Mentioned in more than two sentences before
⋮	⋮

Table 4: Examples of case/location categories.

discourse preferences. Since Web text slightly adheres to formal grammar and thus rich syntactic preferences are not considered very important for resolution of zero anaphora in Web text, we only introduce simple features as non-lexical features.

Case/location We use case/location features to reflect syntactic, functional, and locational preferences. We considered 85 case/location categories, examples of which are summarized in Table 4. If an antecedent candidate appears in a certain case/location category, the corresponding feature value is 1; otherwise 0. These features are made for each case, respectively, i.e. there are a total of 255 case/location features.

Salience Previous work has reported the usefulness of salience in anaphora resolution (Lappin and Leass, 1994; Mitkov et al., 2002; Sasano et al., 2008). We introduce salience features to take into account the salience of each discourse entity. First, we apply the following simple rules to estimating the salience of each entity, and we then set salience features, whose value is 1 if the salience of an antecedent candidate is no less than 1.0; otherwise 0.

- $+2$: mentioned with topical marker "*wa*."
- $+1$: mentioned without topical marker "*wa*."
- $\times 0.5$: beginning of each sentence.

Case assigned features We introduce case assigned features for each case type. The value is 1 if the corresponding zero pronoun is assigned to an antecedent; otherwise 0.

If the weights for these features become larger, corresponding zero pronouns are assigned to antecedents more often. Thus, the weights for these features are regarded as parameters to control the recall/precision trade-off. Although we mainly evaluate our system by using the F-measure, the algorithm mentioned in Section 4.3 does not select a parameter that maximizes the F-measure. Thus, after parameters are estimated by the algorithm, we adjust the weights for these features to maximize the F-measure by using training or development data.

6 Experiments

6.1 Setting

We used the same data set as described in (Sasano et al., 2009). This data set consisted of 186 Web documents (979 sentences, 19,677 morphemes), in which all predicate-argument relations were manually annotated. There were 2,137 predicates in this corpus, and 683 zero anaphoric relations were annotated. We call this data set a Web Corpus after this. We performed 6-fold cross-validation. We used correct morphemes, named entities, dependency structures, and coreference relations that were manually annotated to concentrate on zero anaphora resolution. We tested 10 initial values as parameter Λ. Since the parameters converged to almost the same value for each initial value, we considered that our model achieved the global maximum parameters in most cases.

We applied two baseline models for comparison. The first was the model proposed by Sasano et al. (2008), which did not use a log-linear framework but exploited almost the same clues. In addition, we also conducted an experiment with merged case frames to verify the usefulness of the

	Recall	Precision	F-measure
Sasano et al. (2008)'s model	0.341 (233/683)	0.306 (233/762)	0.322
Proposed model with merged case frames	0.334 (228/683)	0.412 (228/553)	0.369
Proposed model	0.379 (259/683)	0.403 (259/642)	**0.391**

Table 5: Experimental results of zero anaphora resolution on Web Corpus.

Case	Type	Recall	Precision	F-measure
NOM	Intra-sentential	0.504 (120/238)	0.460 (120/261)	0.481
	Inter-sentential	0.460 (104/226)	0.387 (104/269)	0.420
ACC	Intra-sentential	0.250 (17/68)	0.447 (17/38)	0.321
	Inter-sentential	0.163 (7/43)	0.194 (7/36)	0.177
DAT	Intra-sentential	0.105 (6/57)	0.316 (6/19)	0.158
	Inter-sentential	0.098 (5/51)	0.263 (5/19)	0.143

Table 6: Detailed experimental results for the proposed model on the Web Corpus.

Removed feature type	Recall	Precision	F-measure
None (Use all features)	0.379 (259/683)	0.403 (259/642)	**0.391**
(Lexical features)			
− Word PMI	0.363 (248/683)	0.378 (248/656)	*0.370*
− Cluster PMI	0.375 (256/683)	0.401 (256/638)	0.387
− Category PMI	0.367 (251/683)	0.423 (251/593)	**0.393**
− NE PMI	0.381 (260/683)	0.389 (260/668)	0.385
− GW PMI (Clust. + Cat. + NE)	0.350 (239/683)	0.413 (239/579)	*0.379*
− All PMI (Word + GW)	0.325 (222/683)	0.391 (222/567)	*0.355*
− Occupancy of case slot	0.365 (249/683)	0.404 (249/617)	0.383
− Overt argument assignment score	0.411 (281/683)	0.350 (281/804)	*0.378*
(Non-lexical features)			
− Case/location	0.264 (180/683)	0.346 (180/520)	*0.299*
− Salience	0.376 (257/683)	0.411 (257/626)	**0.393**
− All non-lexical (Case/loc. + Salience)	0.250 (171/683)	0.312 (171/545)	*0.279*

Table 7: Performance by removing one feature type at a time from feature sets. Bold font indicates higher F-measures than 0.390 and italics indicate F-measures lower than 0.380.

case frames that had been constructed for each meaning of each verb/adjective. We merged all case frames of a certain verb/adjective in this experiment, and thus each predicate only had one case frame. As a result, this model only took into account knowledge of the relations between two terms, i.e. a predicate and its particular argument. For example, while the model with the case frames in Table 1 considers the dative case of '*te-wo yaku*' would be filled by a *person*, the model with the merged case frame considers the case would be filled by a *disk* with high probability.

6.2 Experimental Results

Table 5 summarizes the experimental results of zero anaphora resolution on the Web Corpus. The results indicate that our proposed model outperformed both Sasano et al. (2008)'s model[7] and the model with merged case frames, which demonstrates the effectiveness of the log-linear framework and the usefulness of the case frames that were constructed for each meaning of each verb/adjective.

Table 6 shows the performance of the proposed model for each case and for each of intra- and inter-sentential zero anaphoric relations. Since the Web Corpus consists of relatively short sentences, there are many inter-sentential zero anaphora. Compared with previous work (Taira et al., 2008; Imamura et al., 2009), our model can resolve inter-sentential zero anaphora in the Web Corpus with comparatively good performance.

6.3 Contribution of features

We eliminated feature types one by one to investigate the contribution each made. Table 7 presents the eliminated feature types and the performance without each type. This table indicates the importance of word PMI and case/location features, since we obtained 0.021 and 0.092 lower F-measures without these features, respectively.

On the other hand, generalized example versions of word PMI did not affect performance much. However, when all generalized example PMIs were eliminated, performance worsened. Therefore, we considered that cluster PMI, category PMI, and NE PMI could be clues for zero anaphora resolution, and confirmed that zero anaphora resolution could benefit from overlap-

[7] For the same 20 articles that were used for testing in (Sasano et al., 2008), our model achieved a recall of 0.525 (64/122), a precision of 0.615 (64/104), and an F-measure of 0.566, while they obtained a recall of 0.410 (50/122), a precision of 0.373 (50/134), and an F-measure of 0.391.

Feature	Weight		
	ga nominative	*wo* accusative	*ni* dative
Occupancy of case slot	0.292	**0.531**	**0.723**
Word PMI	0.154	**0.299**	0.211
Cluster PMI	0.005	**0.347**	0.058
Category PMI	**0.844**	**0.617**	0.391
NE PMI	**0.563**	-0.119	-0.444

Table 8: Weights of lexical features. The Bold font indicates that value of weight is larger than average of weights in the same row.

ping features. Salience features did not contribute to performance when using case/location features. We think this is because case/location features involve salience information.

Table 8 lists the weights of the lexical features in the proposed model. Since the variance in each type of feature is different, it is not very meaningful to compare different feature types. However, we can find a tendency for each case type by comparing the weights of the same feature types. In fact, we have obtained several interesting findings.

The weights of the case slot occupancy features, which denote how often the target slot is occupied by an overt argument, are large for the *wo* (accusative) and *ni* (dative) cases, but small for the *ga* (nominative) case. This means that there are tight relations between occupancy of the case slot and the generativity of the slot in the accusative and dative cases, but not in the case of the nominative.

We also found differences between the nominative and the accusative cases in the PMI features. The weights of course-grained lexical knowledge, such as category and NE PMIs, are relatively large in the nominative case. However, the weights of fine-grained lexical knowledge are relatively large in the accusative case. This means that the lexical preference for the nominative is coarser than that for the accusative case.

6.4 Comparison with previous work

We also conducted experiments on the NAIST Text Corpus version 1.4β (Iida et al., 2007) to compare our results with those from previous work. We used articles from January 1st to 11th and editorials from January to August for training, articles on January 12th and 13th and the September editorials for development, and articles from January 14th to 17th and editorials from October to December for testing. While the NAIST Text Corpus has the predicate-argument structure of the original form annotated even for predicates that appear in passive or causative voice, our model outputs the surface predicate-argument structure. Therefore, we excluded such predicates. Table 9 summarizes the performance of our model on the NAIST Text Corpus with the performance of Imamura et al. (2009)'s model. Although these experiments did not use the exact same data, we can see that our model achieved comparable performance even for newspaper articles.

Case	Type	Imamura et al. (2009) R	P	F	Our model R	P	F
NOM	Intra	0.434	0.588	**0.500**	0.400	0.390	0.395
	Inter	0.076	0.475	0.131	0.221	0.273	**0.244**
ACC	Intra	0.216	0.537	**0.308**	0.169	0.181	0.175
	Inter	0.004	0.250	0.007	0.050	0.101	**0.066**
DAT	Intra	0.000	0.000	0.000	0.098	0.082	**0.089**
	Inter	0.000	0.000	0.000	0.030	0.023	**0.026**

Table 9: Experimental results on the NAIST Text Corpus. R, P, and F denote recall, precision, and F-measure, respectively.

Compared with Iida et al. (2006)'s model, which exploited rich syntactic patterns, our model did not seem to perform as well for the NAIST Text Corpus.[8] We conjectured that this was because the NAIST Text Corpus consisted of newspaper articles that included relatively long sentences with formal grammar, and thus rich syntactic patterns were quite effective. Our model also took into account syntactic clues by using case/location features. However, since we basically focused on the zero anaphora in Web text that only adhered slightly to formal grammar, we did not give priority to exploring effective syntactic patterns.

7 Conclusion

This paper presented a discriminative model for Japanese zero anaphora resolution that can exploit large-scale lexicalized case frames, as well as non-lexical features obtained from a relatively small annotated corpus. Experimental results on a Web text revealed that our model could effectively resolve zero anaphora. We plan to investigate new features for zero anaphora resolution including richer syntactic patterns and global constraints in future work.

[8]Note that, their experiment was conducted on a small and relatively reliable subset of the NAIST Text Corpus, and thus we could not directly compare the results.

References

A. Ferrandez and J. Peral. 2000. A computational approach to zero-pronouns in spanish. In *Proc. of ACL'00*, pages 166–172.

C. J. Fillmore, C. R. Johnson, and M. R.L. Petruck. 2003. Background to framenet. *International Journal of Lexicography*, 16(3):235–250.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Proc. of NAACL-HLT'10*, pages 385–393.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2006. Exploiting syntactic patterns as clues in zero-anaphora resolution. In *Proc. of COLING-ACL'06*, pages 625–632.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a japanese text corpus with predicate-argument and coreference relations. In *Proc. of ACL'07 Workshop: Linguistic Annotation Workshop*, pages 132–139.

Kenji Imamura, Kuniko Saito, and Tomoko Izumi. 2009. Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution. In *Proc. of ACL-IJCNLP'09*, pages 85–88.

IREX Committee, editor. 1999. *Proc. of the IREX Workshop*.

Daisuke Kawahara and Sadao Kurohashi. 2002. Fertilization of case frame dictionary for robust Japanese case analysis. In *Proc. of COLING'02*, pages 425–431.

Daisuke Kawahara and Sadao Kurohashi. 2004. Zero pronoun resolution based on automatically constructed case frames and structural preference of antecedents. In *Proc. of IJCNLP'04*, pages 334–341.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In *Proc. of HLT-NAACL'06*, pages 176–183.

Jun'ichi Kazama and Kentaro Torisawa. 2008. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proc. of ACL-HLT'08*, pages 407–415.

Fang Kong and Guodong Zhou. 2010. A tree kernel-based unified framework for chinese zero anaphora resolution. In *Proc. of EMNLP'10*, pages 882–891.

Anna Korhonen, Yuval Krymolowski, and Ted Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. In *Proc.of LREC'06*, pages 3000–3006.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–562.

Ruslan Mitkov, Richard Evans, and Constantin Orăsan. 2002. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In *Proc. of CICLing'02*, pages 168–186.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proc. of ACL'02*, pages 104–111.

Jorge Nocedal. 1980. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1):71–105.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2008. A fully-lexicalized probabilistic model for japanese zero anaphora resolution. In *Proc. of COLING'08*, pages 769–776.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2009. The effect of corpus size on case frame acquisition for discourse analysis. In *Proc. of NAACL-HLT'09*, pages 521–529.

Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. 2002. A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In *Proc. of COLING'02*, pages 911–917.

Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. 2008. A japanese predicate argument structure analysis using decision lists. In *Proc. of EMNLP'08*, pages 523–532.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and resolution of Chinese zero pronouns: A machine learning approach. In *Proc. of EMNLP-CoNLL'07*, pages 541–550.