Japanese Zero Reference Resolution Considering Exophora and Author/Reader Mentions

Masatsugu Hangyo **Daisuke Kawahara** **Sadao Kurohashi**
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku
Kyoto, 606-8501, Japan
{hangyo,dk,kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

In Japanese, zero references often occur and many of them are categorized into zero exophora, in which a referent is not mentioned in the document. However, previous studies have focused on only zero endophora, in which a referent explicitly appears. We present a zero reference resolution model considering zero exophora and author/reader of a document. To deal with zero exophora, our model adds pseudo entities corresponding to zero exophora to candidate referents of zero pronouns. In addition, we automatically detect mentions that refer to the author and reader of a document by using lexico-syntactic patterns. We represent their particular behavior in a discourse as a feature vector of a machine learning model. The experimental results demonstrate the effectiveness of our model for not only zero exophora but also zero endophora.

1 Introduction

Zero reference resolution is the task of detecting and identifying omitted arguments of a predicate. Since the arguments are often omitted in Japanese, zero reference resolution is essential in a wide range of Japanese natural language processing (NLP) applications such as information retrieval and machine translation.

(1)	パスタが	好きで	毎日	(ϕ ガ)
	pasta-NOM	like	everyday	(ϕ-NOM)
	(ϕ ヲ)	食べます。		
	(ϕ-ACC)	eat		

(Liking pasta, (ϕ) eats (ϕ) every day)

For example, in example (1) , the accusative argument of the predicate "食べます" (eat) is omitted .[1] The omitted argument is called a zero pronoun. In this example, the zero pronoun refers to "パスタ" (pasta).

Zero reference resolution is divided into two subtasks: zero pronoun detection and referent identification. Zero pronoun detection is the task that detects omitted zero pronouns from a document. In example (1), this task detects that there are the zero pronouns in the accusative and nominative cases of "食べます" (eat) and there is no zero pronoun in the dative case of "食べます". Referent identification is the task that identifies the referent of a zero pronoun. In example (1), this task identifies that the referent of the zero pronoun in the accusative case of "食べます" is "パスタ" (pasta). These two subtasks are often resolved simultaneously and our proposed model is a unified model.

Many previous studies (Imamura et al., 2009; Sasano et al., 2008; Sasano and Kurohashi, 2011) have treated only **zero endophora**, which is a phenomenon that a referent is mentioned in a document, such as "パスタ" (pasta) in example (1). However, **zero exophora**, which is a phenomenon that a referent does not appear in a document, often occurs in Japanese when a referent is an author or reader of a document or an indefinite pronoun. For example, in example (1), the referent of the zero pronoun of the nominative case of "食べます" (eat) is the author of

[1] In this paper, we use the following abbreviations: NOM (nominative), ABL(ablative), ACC (accusative), DAT (dative), ALL (allative), GEN (genitive), CMI (comitative), CNJ (conjunction), INS(instrumental) and TOP (topic marker).

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 924–934,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

	Zero pronoun	Referent in the document	Example
Zero endophora	Exist	Exist	僕はカフェが好きで毎日 (カフェ ニ)通っている。 (I like cafes and go (to a cafe) everyday.)
Zero exophora	Exist	Not exist	私がメリットを ([*reader*] ニ) 説明させていただきます。 (I would like to explain the advantage (to [*reader*]).)
No zero reference	Not exist	Not exist	あなたはリラックスタイムが (×ニ)過ごせる。 (You can have a relaxing time.) *There is no dative case.

Table 1: Examples of zero endophora, zero exophora and no zero reference.

the document, but the author is not mentioned explicitly.

(2)

最近は	パソコンで	動画を
recently	PC-INS	movie-ACC

([*unspecified:person*] ガ)	見れる。
([*unspecified:person*]-NOM)	can see

(Recently, (people) can see movies by a PC.)

Similarly, in example (2), the referent of the zero pronoun of the nominative case of "見れる" (can see) is an unspecified person.[2]

Most previous studies have neglected zero exophora, as though a zero pronoun does not exist in a sentence. However, such a rough approximation has impeded the zero reference resolution research. In Table 1, in "zero exophora," the dative case of the predicate has the zero pronoun, but in "no zero reference," the dative case of the predicate does not have a zero pronoun. Treating them with no distinction causes a decrease in accuracy of machine learning-based zero pronoun detection due to a gap between the valency of a predicate and observed arguments of the predicate. In this work, to deal with zero exophora explicitly, we provide pseudo entities such as [*author*], [*reader*] and [*unspecified:person*] as candidate referents of zero pronouns.

In the referent identification, selectional preferences of a predicate (Sasano et al., 2008; Sasano and Kurohashi, 2011) and contextual information (Iida et al., 2006) have been widely used. The author and reader (A/R) of a document have not been used for contextual clues because the A/R rarely appear in the discourse in corpora based on newspaper articles, which are main targets of the previous studies. However, in other domain documents such as blog articles and shopping sites, the A/R often appear in the discourse. The A/R tend to be omitted and there are many clues for the referent identification about the A/R such as honorific expressions and modality expressions. Therefore, it is important to deal with the A/R of a document explicitly for the referent identification.

The A/R appear as not only the exophora but also the endophora.

(3)

僕 $_{author}$ は	京都に	(僕ガ)
I-TOP	Kyoto-DAT	(I-NOM)

行こうと	思っています。
will go	have thought

(I have thought (I) will go to Kyoto.)

皆さん $_{reader}$ は	どこに	行きたいか
you all-TOP	where-DAT	want to go

(皆さんガ)	(僕ニ)	教えてください。
(you all-NOM)	(I-DAT)	let me know

(Please let (me) know where do you want to go.)

In example (3), "僕" (I), which is explicitly mentioned in the document, is the author of the document and "皆さん" (you all) is the reader. In this paper, we call these expressions, which refer to the author and reader, **author mention** and **reader mention**. We treat them explicitly to improve the performance of zero reference resolution. Since the A/R are mentioned as various expressions besides personal pronouns in Japanese, it is difficult to detect the A/R mentions based merely on lexical information. In this work, we automatically detect the A/R mentions by using a learning-to-rank algorithm(Herbrich et al., 1998; Joachims, 2002) that uses lexico-syntactic patterns as features.

Once the A/R mentions can be detected, their information is useful for the referent identification.

[2] In the following examples, omitted arguments are put in parentheses and exophoric referents are put in square brackets.

The A/R mentions have both a property of the discourse element mentioned in a document and a property of the zero exophoric A/R. In the first sentence of example (3), it can be estimated that the referent of the zero pronoun of the nominative case of “行こう” (will go) from a contextual clue that “僕” (I) is the topic of this sentence and a syntactic clues that “僕” (I) depends on “思っています” (have thought) over the predicate “行こう” (will go).[3] Such contextual clues can be available only for the discourse entities that are mentioned explicitly. On the other hand, in the second sentence, since “教えてください” (let me know) is a request form, it can be assumed that the referent of the zero pronoun of the nominative case is “僕” (I), which is the author, and the one of the dative case is “皆様” (you all), which is the reader. The clues such as request forms, honorific expressions and modality expressions are available for the author and reader. In this work, to represent such aspect of the A/R mentions, both the endophora and exophora features are given to them.

In this paper, we propose a zero reference resolution model considering the zero exophora and the author/reader mentions, which resolves the zero reference as a part of a predicate-argument structure analysis.

2 Related Work

Several approaches to Japanese zero reference resolution have been proposed.

Iida et al. (2006) proposed a zero reference resolution model that uses the syntactic relations between a zero pronoun and a candidate referent as a feature. They deal with zero exophora by judging that a zero pronoun does not have anaphoricity. However, the information of zero pronoun existences is given and thus they did not address zero pronoun detection.

Zero reference resolution has been tackled as a part of predicate-argument structure analysis. Imamura et al. (2009) proposed a predicate-argument structure analysis model based on a log-linear model that simultaneously conducts zero endophora resolution. They assumed a particular candidate referent, NULL, and when the analyzer selected this referent, the analyzer outputs “zero exophora or no zero pronoun,” in which they are treated without distinction. Sasano et al. (2008) proposed a probabilistic predicate-argument structure analysis model including zero endophora resolution by using wide-coverage case frames constructed from a web corpus. Sasano and Kurohashi (2011) extended the Sasano et al. (2008)’s model by focusing on zero endophora. Their model is based on a log-linear model that uses case frame information and the location of a candidate referent as features. In their work, zero exophora is not treated and they assumed that a zero pronoun is absent when there is no referent in a document.

For languages other than Japanese, zero pronoun resolution methods have been proposed for Chinese, Portuguese, Spanish and other languages. In Chinese, Kong and Zhou (2010) proposed tree-kernel based models for three subtasks: zero pronoun detection, anaphoricity decision and referent selection. In Portuguese and Spanish, only a subject word is omitted and zero pronoun resolution has been tackled as a part of coreference resolution. Poesio et al. (2010) and Rello et al. (2012) detected omitted subjects and made a decision whether the omitted subject has anaphoricity or not as preprocessing of coreference resolution systems.

3 Baseline Model

In this section, we describe a baseline zero reference resolution system. In our model, the zero reference resolution is conducted as a part of predicate-argument structure (PAS) analysis. The PAS consists of a case frame and an alignment between case slots and referents. The case frames are constructed for each meaning of a predicate. Each case frame describes surface cases that each predicate has (case slot) and words that can fill each case slot (example). In this study, the case frames are constructed from 6.9 billion Web sentences by using Kawahara and Kurohashi (2006a)’s method.

The baseline model does not treat zero exophora as the previous studies. The baseline model analyzes a document in the following procedure in the same way as the previous study (Sasano and Kurohashi, 2011).[4]

[3] Since “僕” (I) depends on “思っています” (have thought), the relation between “僕” (I) and “行こう” (will go) is the zero reference.

[4] For learning, the previous study used a log-linear model, but we use a learning-to-rank model. In our preliminary exper-

京都駅に	ある	カレー屋が	好きで、	その店に	よく	行きます。
Kyoto station-DAT	stand	curry shop-NOM	like	the shop	often	go

(I like a curry shop in Kyoto station and often go to the shop.)

今日は	皆さんに	(カレー屋ヲ)	紹介します。
Today-TOP	you all-DAT	(curry shop-ACC)	will introduce

(Today, I will introduce (the shop) to you.)

Discourse entities

{ 京都駅 (Kyoto station)}, { カレー屋 (curry shop), その店 (the shop)}, { 今日 (today)},
{ 皆さん (you all)}

Candidate predicate-argument structures of "紹介します" in the baseline model

[1-1] case frame:[紹介する (1)], { NOM:Null, ACC:Null, DAT:皆さん, TIME:今日 }
[1-2] case frame:[紹介する (1)], { NOM:Null, ACC:カレー屋, DAT:皆さん, TIME:今日 }
[1-3] case frame:[紹介する (1)], { NOM:京都駅, ACC:カレー屋, DAT:皆さん, TIME:今日 }
⋮
[2-1] case frame:[紹介する (2)], { NOM:Null, ACC:Null, DAT:皆さん, TIME:今日 }
[2-2] case frame:[紹介する (2)], { NOM:Null, ACC:カレー屋, DAT:皆さん, TIME:今日 }
⋮

Figure 1: Examples of discourse entities and predicate-argument structures

1. Parse the input document and recognize named entities.
2. Resolve coreferential relations and set discourse entities.
3. Analyze the predicate-argument structure for each predicate using the following steps:
   (a) Generate candidate predicate-argument structures.
   (b) Calculate the score of each predicate-argument structure and select the one with the highest score.

We illustrate the details of the above procedure. First, we describe how to set the discourse entities in step 2. In our model, we treat referents of a zero pronoun using a unit called **discourse entity**, which is what mentions in a coreference chain are bound into. In Figure 1, we treat "カレー屋" (curry shop) and "その店" (the shop), which are in a coreference chain, as one discourse entity. In Figure 1, the discourse entity { カレー屋, その店 } is selected for the referent of the accusative case of the predicate " 紹介します" (will introduce).

Next, we illustrate the PAS analysis in step 3. In step 3a, possible combinations of the case frame (cf) and the alignment (a) between case slots and discourse entities are listed. First, one case frame is selected from case frames for the predicate. Next, overt arguments, which have dependency relations with the predicate, are aligned to a case slot of the case frame. Finally, each of zero pronouns of remaining case slots is assigned to a discourse entity or is not assigned to any discourse entities. The case slot whose zero pronoun is not assigned to any discourse entities corresponds to the case that does not have a zero pronoun. In Figure 1, we show the examples of candidate PASs. In these examples, [紹介する (1)] and [紹介する (2)] are case frames corresponding to each meaning of "紹介する". Referents of each case slot are actually selected from discourse entities but are explained as a representative word for illustration. "Null" indicates that a case slot is not assigned to any discourse entities. Since alignments between case slots and discourse entities of the PAS [1-2] and [2-2] are the same but their case frames are different, we deal with them as discrete PASs. In this case, however, the results of zero reference resolution are the same.

We represent each PAS as a feature vector, which is described in section 3.1, and calculate a score of each PAS with the learned weights. Finally, the system outputs the PAS with the highest score.

iment of the baseline model, there is little difference between the results of these methods.

Type	Value	Description
Case frame	Log	Probabilities that {words, categories and named entity types} of e is assigned to c of cf
	Log	Generative probabilities of {words, categories and named entity types} of e
	Log	PMIs between {words, categories and named entity types} of e and c of cf
	Log	Max of PMIs between {words, categories and named entity types} of e and c of cf
	Log	Probability that c of cf is assigned to any words
	Log	Ratio of examples of c to ones of cf
	Binary	c of cf is {adjacent and obligate} case
Predicate	Binary	Modality types of p
	Binary	Honorific expressions of p
	Binary	Tenses of p
	Binary	p is potential form
	Binary	Modifier of p (predicate, noun and end of sentence)
	Binary	p is {dynamic and stative} verb
Context	Binary	Named entity types of e
	Integer	Number of mentions about e in t
	Integer	Number of mentions about e {before and after} p in t
	Binary	e is mentioned with post position "は" in a target sentence
	Binary	Sentence distances between e and p
	Binary	Location categories of e (Sasano and Kurohashi, 2011)
	Binary	e is mentioned at head of a target sentence
	Binary	e is mentioned with post position {"は" and "が" } at head of a target sentence
	Binary	e is mentioned at head of the first sentence
	Binary	e is mentioned with post position "は" at head of the first sentence
	Binary	e is mentioned at end of the first sentence
	Binary	e is mentioned with copula at end of the first sentence
	Binary	e is mentioned with noun phrase stop at end of the first sentence
	Binary	Salience score of e is larger than 1 (Sasano and Kurohashi, 2011)
other	Binary	c is assigned

Table 2: The features of $\phi_{assigned}(cf, c \leftarrow e, p, t)$

3.1 Feature Representation of Predicate-Argument Structure

When text t and target predicate p are given and PAS (cf, a) is chosen, we represent a feature vector of the PAS as $\phi(cf, a, p, t)$. $\phi(cf, a, p, t)$ consists of a feature vector $\phi_{overt\text{-}PAS}(cf, a, p, t)$ and feature vectors $\phi(cf, c/e, p, t)$. Where $\phi_{overt\text{-}PAS}(cf, a, p, t)$ corresponds to alignment between case slots and overt (not omitted) arguments and $\phi(cf, c/e, p, t)$ represents that a case slot c is assigned to a discourse entity e. If a case slot is assigned to an overt entity, $\phi(cf, c/e, p, t)$ is set to a zero vector.

Each feature vector $\phi(cf, c/e, p, t)$ consists of $\phi_A(cf, c/e, p, t)$ and $\phi_{NA}(cf, c/Null, p, t)$. $\phi_A(cf, c/e, p, t)$ becomes active when the case slot c is assigned to the discourse entity e and $\phi_{NA}(cf, c/Null, p, t)$ becomes active when the case slot c is not assigned to any discourse entities. For example, the PAS [1-2] in Figure 1 is represented as:
($\phi_{overt\text{-}PAS}$(紹介する (1), {NOM:Null, ACC:Null, NOM:皆さん, TIME:今日 }), $\mathbf{0}_{\phi_A}$,
ϕ_{NA}(紹介する (1), NOM/$Null$),
ϕ_A(紹介する (1), ACC/カレー屋),
$\mathbf{0}_{\phi_{NA}}, \mathbf{0}_{\phi_A}, \mathbf{0}_{\phi_{NA}}$). [5]
In our feature representation, the second and third terms correspond to the nominative case, the forth and fifth ones correspond to the accusative and the sixth and seventh ones correspond to the dative case.

We present the details of $\phi_{overt\text{-}PAS}(cf, a, p, t)$, $\phi_A(cf, c/e, p, t)$ and $\phi_{NA}(cf, c/Null, p, t)$. We use a score of the probabilistic PAS analysis (Kawahara and Kurohashi, 2006b) to $\phi_{overt\text{-}PAS}(cf, a, p, t)$. We list the features of $\phi_A(cf, c/e, p, t)$ in Table 2 and the features of $\phi_{NA}(cf, c/Null, p, t)$ in Table

[5] In the following example, p and t are sometimes omitted and $\mathbf{0}_\phi$is 0 vector that has the same dimension as ϕ.

Type	Value	Description
Case frame	Log	Probability that c of cf is not assigned
	Log	Ratio of number of examples of c to ones of cf
	Binary	c of cf is $\{$adjacent and obligate$\}$ case

Table 3: The features of $\phi_{NA}(cf, c/Null, p, t)$

3.

3.2 Weight Learning

In the previous section, we defined the feature vector $\phi(cf, a, p, t)$, which represents a PAS. In this section, we illustrate the learning method of the weight vector corresponding to the feature vector. The weight vector is learned by using a learning-to-rank algorithm.

In a corpus, gold-standard alignments a^* are manually annotated but case frames are not annotated. Since the case frames are constructed for each meaning, some of them are unsuitable for a usage of a prdicate in a context. If training data includes PASs (cf, a^*) whose cf is such case a frame as correct instances, these are harmful for training. Hence, we treat a case frame cf^* which is selected by a heuristic method as a correct case frame and remove (cf, a^*) which has other cf.

In particular, we make ranking data for the learning for each target predicate p in the following steps.

1. List possible PASs (cf, a) for predicate p.
2. Calculate a probabilistic zero reference resolution score for each (cf, a^*) and define the one with highest score as (cf^*, a^*).
3. Remove (cf, a^*) except (cf^*, a^*) from the learning instance.
4. Make ranking data that (cf^*, a^*) has a higher rank than other (cf, a).

In the above steps, we make ranking data for each predicate and use the ranking data collected from all target predicates as training data.

4 Corpus

In this work, we use Diverse Document Leads Corpus (DDLC) (Hangyo et al., 2012) for experiments. In DDLC, documents collected from the web are annotated with morpheme, syntax, named entity, coreference, PAS and A/R mention. Morpheme, syntax, named entity, coreference and PAS are annotated on the basis of Kyoto University Text Corpus (Kawahara et al., 2002). The PAS annotation includes zero reference information and the exophora referents are defined as five elements, [*author*], [*reader*], [*US(unspecified):person*], [*US:matter*] and [*US:situation*]. The A/R mentions are annotated to head phrases of compound nouns when the A/R mentions consist of compound nouns. If the A/R is mentioned by multiple expressions, only one of them is annotated with the A/R mention tag and all of these mentions are linked by a coreference chain. In other words, the A/R mentions are annotated to discourse entities. In the web site of an organization such as a company, the site administrator often writes the document on behalf of the organization. In such a case, the organization is annotated as the author.

5 Author/Reader Mention Detection

A/R mentions, which refer to A/R of a document, have different properties from other discourse entities. The A/R are mentioned as very various expressions such as personal pronouns, proper expressions and role expressions.

(4) こんにちは、 企画チームの
Hello project team-GEN
梅辻 $_{author}$ です。
am Umetsuji
(Hello, I'm Umetsuji on the project team.)

(5) 問題が あれば 管理人 $_{author}$ まで
problem-NOM exist to moderator
お知らせください。
let me know
(Please let me know if there are any problems.)

In example (4), the author is mentioned as "梅辻" (Umetsuji), which is the name of the author, and in example (5), the author is mentioned as "管理人" (moderator), which expresses the status of the author. Likewise, the reader is sometimes mentioned as "お客様" (customer) and others. However, since such expressions often refer to someone other than the A/R, whether an expression indicates the A/R of a document depends on the context of the document.

In English and other languages, the A/R mentions can be detected from coreference information because it can be assumed that the expression that has a coreference relation with first or second personal pronoun is the A/R mention. However, since the A/R tend to be omitted and personal pronouns are rarely used in Japanese, it is difficult to detect the A/R mentions from coreference information. Because of these reasons, it is difficult to detect which discourse entity is the A/R mention from lexical information of the entities. In this study, the A/R mentions are detected from lexico-syntactic (LS) patterns in the document. We use a learning-to-rank algorithm to detect A/R mentions by using the LS patterns as features.

5.1 Author/Reader Detection Model

We use a learning-to-rank method for detecting A/R mentions. This method learns the ranking that entities of the A/R mentions have a higher rank than other discourse entities. Here, it is an important point that there are no A/R mentions in some documents. The documents in which the A/R mentions do not appear are classified into two types. The first type is a document that the A/R do not appear in the discourse of the document such as newspaper articles and novels. The second type is a document that the A/R appear in the discourse but all of their mentions are omitted. For example, in Figure 1, the author appears in the discourse (e.g. the nominative argument of "like") but is not mentioned explicitly. We introduce two pseudo entities corresponding to these types. The first pseudo entity "no A/R mention (discourse)" represents the document that the A/R do not appear in the discourse. It is considered that the document that the A/R do not appear have characteristics of writing style such that honorific expressions and request expressions are rarely used. This pseudo entity is represented as a document vector that consists of LS pattern features of the whole document, which reflect a writing style of a document. The second pseudo entity "no A/R mention (omitted)" represents the document in which all mentions of the A/R are omitted and this pseudo entity is represented as 0 vector. Since a decision score of this pseudo entity is allways 0, discourse entities whose score is lower than the score of this pseudo entity can be treated as a negative example in a binary classification.

When there are A/R mentions in a document, we make ranking data where the discourse entity of the A/R mention has a higher rank than other discourse entities and "no A/R mention" pseudo entities. When the A/R do not appear in the discourse, we make ranking data where "no A/R mention (discourse)" has a higher rank than all discourse entities and "no A/R mention (omitted)". When the A/R appear in the discourse but all mentions are omitted, we make ranking data where "no A/R mention (omitted)" has a higher rank than all discourse entities and "no A/R mention (discourse)". We judge that the A/R appear in the discourse if the A/R appear as a referent of zero reference in gold-standard PASs and this judgment is used only in the training phase. After making the ranking data for each document, all of the ranking data are merged and the merged data is fed into the learning-to-rank model.

For the A/R mention detection, we calculate the score of all discourse entities and the pseudo entities and select the discourse entity with the highest score to the A/R mention. If any "no A/R mention" have the highest score, we decide that there are no A/R mentions in the document.

5.2 Lexico-Syntactic Patterns

For each discourse entity, phrases of the discourse entity, its parent and their dependency relations are used to make LS patterns that represent the discourse entity. When a discourse entity is mentioned multiple times, the phrases of all mentions are used to make the LS patterns. LS patterns of phrases are made by generalizing these phrases on various levels (types). LS patterns of dependencies are made from combining the LS patterns of phrases.

Table 4 lists generalization types. On the *word* type, we make a phrase LS pattern by generalizing each content word and jointing them. For example, a LS pattern of the phrase "ぼくは" generalized on the <representative form> is "僕は". The *word+* type is the same as *word* except all content words are generalized on the <part of speech and conjugation>. For example, a LS pattern of the dependency relation "太郎は → 走った" generalized on the <named entity> is "NE:PERSON+は → verb:past". We also use the LS patterns of generalized individual morphemes. On the *phrase* type, each phrase is generalized according to the information assigned to the phrase and all content words are generalized on the <part of speech and conjugation> if the information

Unit	Type	Example (original phrase)
word	<no generalization>	僕は (僕は)
	<original form>	走った (走る)
	<representative form>	僕は (ぼくは)
	<part of speech and conjugation>	verb:past (走った)
word+	<category>	Category:PERSON+は (僕は)
	<named entity>	NE:PERSON+は (太郎は)
	<first person pronoun>	FirstPersonPronoun+は (僕は)
	<second person pronoun>	SecondPersonPronoun+に (あなたに)
phrase	<modality>	modality:request (お問い合わせください)
	<honorific expression>	honorific:modest (お送りします)
	<attached words>	ください (お問い合わせください)

Table 4: Generalization types of the LS patterns

is not assigned to the phrase.

For "no A/R mention (discourse)" instance, the above features of all mentions, including verbs and adjectives, and their dependencies in the document are gathered and used as the features representing the instance.

6 Zero Reference Resolution Considering Exophora and Author/Reader Mentions

In this section, we describe the zero reference resolution system that considers the zero exophora and the A/R mentions. The proposed model resolves zero reference as a part of the PAS analysis based on the baseline model.

The proposed model analyzes the PASs in the following steps:

1. Parse the input document and recognize named entities.
2. Resolve coreferential relations and set discourse entities.
3. Detect the A/R mentions of the document.
4. Set pseudo entities from the estimated A/R mentions.
5. Analyze the PAS for each predicate using the same procedure as the baseline model.

The differences form baseline model are the estimation of the A/R mentions in step 3 and the setting of pseudo entities in step 4.

6.1 Pseudo Entities and Author/Reader Mentions for Zero Exophora

In the baseline model, referents of zero pronouns are selected form discourse entities. The proposed model adds pseudo entities([*author*], [*reader*], [*US:person*] (unspecified:person) and [*US:others*] (unspecified:others)[6]) to deal with zero exophora.

When the A/R mentions appear in a document, the A/R pseudo entities raise an issue. The zero endophora are given priority to zero exophora. In other words, the A/R mentions are selected to the referents in preference to pseudo entities when there are A/R mentions. Therefore, when the system estimates that A/R mentions appear, the A/R pseudo entities are not created.

In the PAS analysis, referents are selected from discourse entities and the pseudo entities. A zero reference is the zero exophora when a case slot is assigned to pseudo entities. Candidate PASs of "紹介します" in Figure 1 are shown in Figure 2.

6.2 Feature Representation of Predicate Argument Structure

In the same way as the baseline model, the proposed model represents a PAS as a feature vector that consists of the feature vector $\phi_{overt\text{-}PAS}(cf, a, p, t)$ and the feature vectors $\phi(cf, c/e, p, t)$. The difference from the baseline model is a composition of $\phi_A(cf, c/e, p, t)$. In the proposed model, each $\phi_A(cf, c/e)$ is composed of vectors, $\phi_{discourse}(cf, c/e)$, $\phi_{[author]}(cf, c/e)$, $\phi_{[reader]}(cf, c/e)$, $\phi_{[US:person]}(cf, c/e)$, $\phi_{[US:others]}(cf, c/e)$ and $\phi_{max}(cf, c/e)$. Their contents and dimensions are the same and similar to $\phi_A(cf, c/e)$ of the baseline model the except for the

[6] We merge [*US:matter*] and [*US:situation*] because of the small amount of [*US:situation*] in the corpus.

[1-1] case frame:[紹介する (1)], { NOM:[*author*], ACC:Null, DAT:皆さん $_{reader}$, TIME:今日 }
[1-2] case frame:[紹介する (1)], { NOM:[*US:person*], ACC:Null, DAT:皆さん $_{reader}$, TIME:今日 }
[1-3] case frame:[紹介する (1)], { NOM:[*author*], ACC:カレー屋, DAT:皆さん $_{reader}$, TIME:今日 }
[1-4] case frame:[紹介する (1)], { NOM:京都駅, ACC:カレー屋, DAT:皆さん $_{reader}$, TIME:今日 }
[1-5] case frame:[紹介する (1)], { NOM:[*author*], ACC:[*US:others*], DAT:皆さん $_{reader}$, TIME:今日 }
⋮
[2-1] case frame:[紹介する (2)], { NOM:[*author*], ACC:Null, DAT:皆さん $_{reader}$, TIME:今日 }
[2-2] case frame:[紹介する (2)], { NOM:[*US:person*], ACC:Null, DAT:皆さん $_{reader}$, TIME:今日 }
⋮

Figure 2: Candidate predicate-argument structures of "紹介します" in the proposed model

	Expressions	Categories
author	私 (I), 我々 (we), 俺 (I), 僕 (I), 当社 (our company), 弊社 (our company), 当店 (our shop)	PERSON, ORGANIZATION
reader	あなた (you), 客 (customer), 君 (you), 皆様 (you all), 皆さん (you all), 方 (person), 方々 (people)	PERSON
US:person	人 (person), 人々 (people)	PERSON
US:others	もの (thing), 状況 (situation)	all categories except PERSON and ORGANIZATION

Table 5: Expressions and categories for pseudo entities

addition of a few features described in section 6.3.

$\phi_{discourse}$ corresponds to the discourse entities, which are mentioned explicitly and becomes active when e is a discourse entity including the A/R mentions. $\phi_{discourse}$ is the same as ϕ_A of the baseline model and the difference is explained in section 6.3. $\phi_{[author]}$ and $\phi_{[reader]}$ become active when e is [*author*]/[*reader*] or the discourse entity corresponding to the A/R mention. In particular, when e is the discourse entity corresponding to the A/R mention, both $\phi_{discourse}$ and $\phi_{[author]}$/$\phi_{[reader]}$ become active. This representation gives the A/R mentions the properties of the discourse entity and the A/R. $\phi_{[US:person]}$ and $\phi_{[US:others]}$ become active when e is [*US:person*] and [*US:others*].

Because $\phi_{[author]}$, $\phi_{[reader]}$, $\phi_{[US:person]}$ and $\phi_{[US:others]}$ correspond to the pseudo entities, which are not mentioned explicitly, we cannot use word information such as expressions and categories. We assume that the pseudo entities have expressions and categories shown in Table 5 and use these to calculate case frame features. Finally, ϕ_{max} consists of the highest value of correspondent feature of the above feature vectors.

6.3 Author/Reader Mention Score

We add A/R mention score features to the feature vector $\phi_A(cf, c/e, p, t)$ described in Table 2. The A/R mention scores are the discriminant function scores of the A/R mention detection. When e is the A/R mention, we set the A/R mention score to the feature.

7 Experiments

7.1 Experimental Settings

We used 1,000 documents from DDLC and performed 5-fold cross-validation. 1,440 zero endophora and 1,935 zero exophora are annotated in these documents. 258 documens are annotated with author mentions and 105 documens are annotated with reader mentions. We used gold-standard (manually annotated) morphemes, named entities, dependency structures and coreference relations to focus on the A/R detection and the zero reference resolution. We used SVM^{rank}[7] for the learning-to-rank method of the A/R detection and the PAS analysis. The categories of words are given by the morphological analyzer JUMAN[8]. Named entities and predicate features (e.g., honorific expressions, modality)

[7] http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

[8] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

		System output		
		Exist		None
		Correct	Wrong	
Gold	Exist	140	6	112
-standard	None	-	38	704

Table 6: Result of the author mention detection

		System output		
		Exist		None
		Correct	Wrong	
Gold	Exist	56	2	47
-standard	None	-	23	872

Table 7: Result of the reader mention detection

	Recall	Precision	F1
Baseline	0.269	0.377	0.314
Proposed model (estimate)	0.282	0.448	0.346
Proposed model (gold-standard)	0.388	0.522	0.445

Table 8: Results of zero endophora resolution

	Recall	Precision	F1
Baseline	0.115	0.377	0.176
Proposed model (estimate)	0.317	0.411	0.358
Proposed model (gold-standard)	0.377	0.485	0.424

Table 9: Results of zero reference resolution

are given by the syntactic parser KNP.[9]

7.2 Results of Author/Reader Mention Detection

We show the results of the author and reader mention detection in Table 6 and Table 7. In these tables, “exist” indicates numbers of documents in which the A/R mentions are manually annotated or our system estimated that some discourse entities are A/R mentions. From these results, the A/R mentions including “none” can be predicted to accuracies of approximately 80%. On the other hand, the recalls are not particularly high: the recall of author is 140/258 and the recall of reader is 56/105. This is because the documents in which the A/R do not appear are more than the ones in which the A/R appear and the system prefers to output “no author/reader mention” as the result of training.

7.3 Results of Zero Reference Resolution

We show the results of zero reference resolution in Table 8 and Table 9. The difference between the baseline and the proposed model is statistically significant ($p < 0.05$) from the McNemar’s test. In Table 8, we evaluate only the zero endophora for comparison to the baseline model, which deals with only the zero endophora. “Proposed model (estimate)” shows the result of the proposed model which estimated the A/R mentions and “Proposed model (gold-standard)” shows the result of the proposed model which is given the A/R mentions of gold-standard from the corpus.

From Table 8, considering the zero exophora and the A/R mentions improves accuracy of zero endophora resolution as well as zero reference resolution including zero exophora.

From Table 8 and Table 9, the proposed model given the gold-standard A/R mentions achieves extraordinarily high accuracies. This result indicates that improvement of the A/R mention detection improves the accuracy of zero reference resolution in the proposed model.

8 Conclusion

This paper presented a zero reference resolution model considering exophora and author/reader mentions. In the experiments, our proposed model achieves higher accuracy than the baseline model. As future work, we plan to improve the author/reader detection model to improve the zero reference resolution.

References

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2012. Building a diverse document leads corpus annotated with semantic relations. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 535–544, Bali,Indonesia, November. Faculty of Computer Science, Universitas Indonesia.

Ralf Herbrich, Thore Graepel, Peter Bollmann-Sdorra, and Klaus Obermayer. 1998. Learning preference relations for information retrieval. In *ICML-98 Workshop: text categorization and machine learning*, pages 80–84.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2006. Exploiting syntactic patterns as clues in zero-anaphora

[9] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 625–632, Sydney, Australia, July. Association for Computational Linguistics.

Kenji Imamura, Kuniko Saito, and Tomoko Izumi. 2009. Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 85–88, Suntec, Singapore, August. Association for Computational Linguistics.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM.

Daisuke Kawahara and Sadao Kurohashi. 2006a. Case frame compilation from the web using high-performance computing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1344–1347.

Daisuke Kawahara and Sadao Kurohashi. 2006b. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 176–183, New York City, USA, June. Association for Computational Linguistics.

Daisuke Kawahara, Sadao Kurohashi, and Koiti Hasida. 2002. Construction of a japanese relevance-tagged corpus. In *Proc. of The Third International Conference on Language Resources Evaluation*, May.

Fang Kong and Guodong Zhou. 2010. A tree kernel-based unified framework for chinese zero anaphora resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 882–891, Cambridge, MA, October. Association for Computational Linguistics.

Massimo Poesio, Olga Uryupina, and Yannick Versley. 2010. Creating a coreference resolution system for italian. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Luz Rello, Ricardo Baeza-Yates, and Ruslan Mitkov. 2012. Elliphant: Improved automatic detection of zero subjects and impersonal constructions in spanish. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 706–715. Association for Computational Linguistics.

Ryohei Sasano and Sadao Kurohashi. 2011. A discriminative approach to japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 758–766, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2008. A fully-lexicalized probabilistic model for japanese zero anaphora resolution. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 769–776, Manchester, UK, August. Coling 2008 Organizing Committee.