Parameter estimation of Japanese predicate argument structure analysis model using eye gaze information

Ryosuke Maki **Hitoshi Nishikawa** **Takenobu Tokunaga**
Department of Computer Science
Tokyo Institute of Technology
{maki.r.aa@m, {hitoshi, take}@c}.titech.ac.jp

Abstract

In this paper, we propose utilising eye gaze information for estimating parameters of a Japanese predicate argument structure (PAS) analysis model. We employ not only linguistic information in the text, but also the information of annotator eye gaze during their annotation process. We hypothesise that annotator's frequent looks at certain candidates imply their plausibility of being the argument of the predicate. Based on this hypothesis, we consider annotator eye gaze for estimating the model parameters of the PAS analysis. The evaluation experiment showed that introducing eye gaze information increased the accuracy of the PAS analysis by 0.05 compared with the conventional methods.

1 Introduction

In recent years, there have been many attempts of annotating corpora with various kinds of information, as supervised machine learning (ML) techniques had been a significant device for natural language processing (NLP) (Pustejovsky and Stubbs, 2012). The annotated corpus is used as training data for constructing a task model, considering the annotated information as expected output of the model. In the current framework, however, only annotated information in the corpus is used for training. In this paper, we propose utilising the information of annotator behaviour during the annotation process as well as resulting annotated information for training the task model (Tokunaga et al., 2013).

We take the predicate argument structure (PAS) analysis of Japanese texts as the target task in the present work. Predicate argument relations are usually marked by case particles denoting grammatical cases in Japanese, therefore identifying dependencies marked by the major obligatory cases, *ga* (nominative), *wo* (accusative) and *ni* (dative) is the main task. However, since ellipses are ubiquitous in Japanese texts, arguments might be identified beyond the sentence including the target predicates (inter-sentence arguments) as well as within the sentence (intra-sentence arguments). This feature is different from the PAS analysis in English in which arguments can be found in the same sentence of the target predicate in most cases. Another complication is the case alternation caused by certain types of auxiliary verbs such as causative and passivisation verbs. In such cases, the original case should be recovered in the PAS analysis. In this respect, the Japanese PAS analysis shares the similarity with semantic role labelling in English (Gildea and Jurafsky, 2002). Furthermore, treating "event nouns" (Komachi et al., 2007) as predicates, we identify their arguments as well as the arguments of verbs and adjectives. Currently, several PAS annotated Japanese corpora such as NAIST Text Corpus (NTC) (Iida et al., 2007b) and BCCWJ-DepParaPAS (Ueda et al., 2015; Maekawa et al., 2014) are available. We use the latter in this study.

In order to improve the PAS analysis performance, we propose to utilise the information of annotator behaviour, particularly eye gaze information, during their annotating predicate argument relations in texts. In the past PAS analysis, a model has been constructed by utilising a certain ML technique regarding the human annotated argument as the correct argument for a given predicate, i.e. it is considered the positive example for the predicate and all other argument candidates are negative examples. Observing

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2861–2869, Osaka, Japan, December 11-17 2016.

the annotator eye movement during their annotation, however, they look at other argument candidates as well until their final decision. Although the not-selected candidates are not the argument of the target predicate in the text, frequently-looked candidates could be an argument of the predicate in similar but different texts. Our idea is to utilise the information gained from those "near miss" candidates in the training phase of the model.

Consider the following two example texts.

(1) *watasi-wa kinou tomodati-to ranti-wo tabeni ikimasita.*
I-TOPIC yesterday friend-with lunch-ACC to eat went

nedan-ni mo azi-ni mo manzokusimasita.
price-DAT also taste-DAT also was satisfied

'Yesterday, I went to have lunch with my friend. I was satisfied with its price and taste.'

(2) *watasi-wa kinou tomodati-to ranti-wo tabeni ikimasita.*
I-TOPIC yesterday friend-with lunch-ACC to eat went

nedan-ni mo azi-ni mo manzokusita youdesu.
price-DAT also taste-DAT also was satisfied seem

'Yesterday, I went to have lunch with my friend. He seemed to be satisfied with its price and taste.'

Although these texts are almost the same on the surface except for the verb ending in the second sentences, the *ga* (nominative) arguments of the predicate '*manzokusuru* (be satisfied)' are different; the *ga* argument in (1) is '*watasi* (I)' in the first sentence while that in (2) is '*tomodati* (friend)'. This difference is caused by the modality auxiliary '*youdesu* (seem)' in the second sentence of (2). Treating these texts as a part of the training examples for binary classification, '*watasi* (I)' is regarded as a positive example, and other candidates including '*tomodati* (friend)' as negative examples in the text (1). However, '*tomodati* (friend)' looks a better negative example than others in (1). This is also the case for (2) but in the opposite way. Our proposal treats '*tomodati* (friend)' in (1) and '*watasi* (I)' in (2) as "near miss" candidates, i.e. better negative examples than others in each text, and takes them into account in the training. To salvage the "near miss" candidates that were merely discarded in the existing PAS analysis based on binary classification, we build our PAS analyser with a learning-to-rank framework; we make a ranking of candidates instead of their binary positive-negative distinction.

It has been reported that the eye gaze information contributes to detecting annotation disagreement between annotators (Mitsuda et al., 2013), but there has been no study on the Japanese PAS analysis using eye gaze information. We hypothesise that frequent looks at argument candidates imply their plausibility of being the argument of the predicate. In this study, we make a candidate ranking based on the frequency of the annotator gaze at the candidates for training a model with the ranking SVM (Joachims, 2002)

This paper is organised as follows. Section 2 overviews the related work, and in Section 3, we define the task setting and propose a method for Japanese PAS analysis. Section 4 discusses the evaluation results and Section 5 concludes the paper.

2 Related work

Basic features for the PAS analysis were proposed in the studies in English (Gildea and Jurafsky, 2002). Unlike English, subject ellipses frequently occur in Japanese. Thus dealing with ellipses, i.e. zero anaphora resolution, is fundamental in processing Japanese texts, and linguistic features for Japanese zero anaphora resolution have been proposed in the past research (Iida et al., 2007a; Sasano and Kurohashi, 2011; Komachi et al., 2007).

The past Japanese PAS analysis methods can be categorised into two classes: one constructs an individual model for predicting each case (*ga* , *wo* , and *ni*) (Taira et al., 2008; Imamura et al., 2009; Hayashibe et al., 2011) and another constructs a single model for predicting all three cases at the same time. The

latter is further divided into two types: (i) identifying all the three cases of one predicate (Sasano and Kurohashi, 2011; Yoshikawa et al., 2011; Hangyo et al., 2013), and (ii) identifying all the three cases of all predicates in the same sentence (Ouchi et al., 2015; Shibata et al., 2016). Since *ga* arguments tend to be omitted more than other cases, we focus on identifying the *ga* case in the present study, thus our proposing method belongs to the former class.

Table 1: Features for PAS analysis

category	ID	feature name	description
predicate	1	lemma	a lemma of the predicate
	2	word origin	originated in Japanese, Chinese, other language or compound
	3	parts of speech	POS of the predicate
	4	conjugation form	a conjugation form of the predicate
	5	conjugation type	a conjugation type of the predicate
	6	surface form	a surface form of the predicate
argument	7	lemma	a lemma of the argument candidate
	8	word origin	originated in Japanese, Chinese, other language or compound
	9	parts of speech	POS of the argument candidate
	10	surface form	a surface form of the argument candidate
	11	case marker	a case particle following the argument candidate
	12	semantic category	a semantic category of the argument candidate
predicate and argument	13	lemma pair	a pair of lemmas of the predicate and the argument candidate
	14	distance	distance between the predicate and the argument candidate
	15	intra/inter	a binary value indicating if the predicate and the argument candidate are in the same sentence or not
	16	intra/inter+case marker	combination of the feature 11 and 15
	17	semantic category pair	a pair of semantic categories of the predicate and the argument candidate
	18	dependency	dependency type: direct dependency, indirect dependency and no dependency

3 Introducing eye gaze information into PAS analysis

3.1 Task setting

The task in this study is identifying the *ga* argument of a specified predicate in a text. We particularly focus on the *ga* case since it tends to be omitted more than other cases in Japanese. Given a target predicate and argument candidates preceding the predicate in a text as an input, the learnt model is expected to select a candidate as the correct *ga* argument as the output. The goal of this study is to show that eye gaze information is useful for training the PAS analysis model.

3.2 Detecting fixations

As detailed below, we detect fixations from the recorded gaze sequence by using the Dispersion-Threshold Identification Algorithm (Salvucci and Goldberg, 2000). A fixation on a word in texts is widely believed to have some relation with the cognitive process on that word (Just and Carpenter, 1980). The overview of the algorithm is described below.

1. A gaze point is added to a set one by one as far as the following conditions hold.
   - All gaze points in the set resides within the distance threshold D from the centre-of-gravity of the set.
   - No tracking error was flagged in the set.

   Just before violating the above conditions, the centre-of-gravity of the set is identified as a fixation candidate.
2. Repeat Step 1. to the rest of the gaze sequence.
3. For each fixation candidate, if its duration time is T or longer, we identify the candidate as a fixation.

Following Mitsuda et al. (2013), we set the space and time threshold D and T as 16 pixels and 100 msec respectively. By making the correspondence between the fixation points and the bounding boxes of argument candidates, we obtain the data indicating what argument candidates the annotators looked at at a certain duration at a certain time point.

3.3 Features

We use the features proposed in the past studies (Taira et al., 2008; Imamura et al., 2009) to represent each argument candidate as shown in Table 1. The distance feature d is calculated by the number of words between the predicate and the candidate, which is normalised to $d \in [0, 1]$. The semantic category feature adopts the categories defined in the Japanese thesaurus 'Nihongo Goi Taikei' (Ikehara et al., 1997).

3.4 Ranking argument candidates

We make the ranking of argument candidates for each training instance represented in terms of the features in Table 1 by using the fixations in the following way.

1. The correct candidate is placed at the top of the ranking.
2. The most frequently fixated candidate other than the correct one follows the correct candidate in the ranking.
3. The candidates with no fixation are placed at the bottom of the ranking with equal rank.
4. Any other candidates are excluded from the ranking.

Using these rankings of argument candidates as training data for the ranking SVM (Joachims, 2002), we estimate the model parameter. In the test phase, the most highly ranked candidate by the model is considered as the model output. Note that we do not require any eye gaze information in the test phase. The eye gaze information is required only for estimating the model parameters.

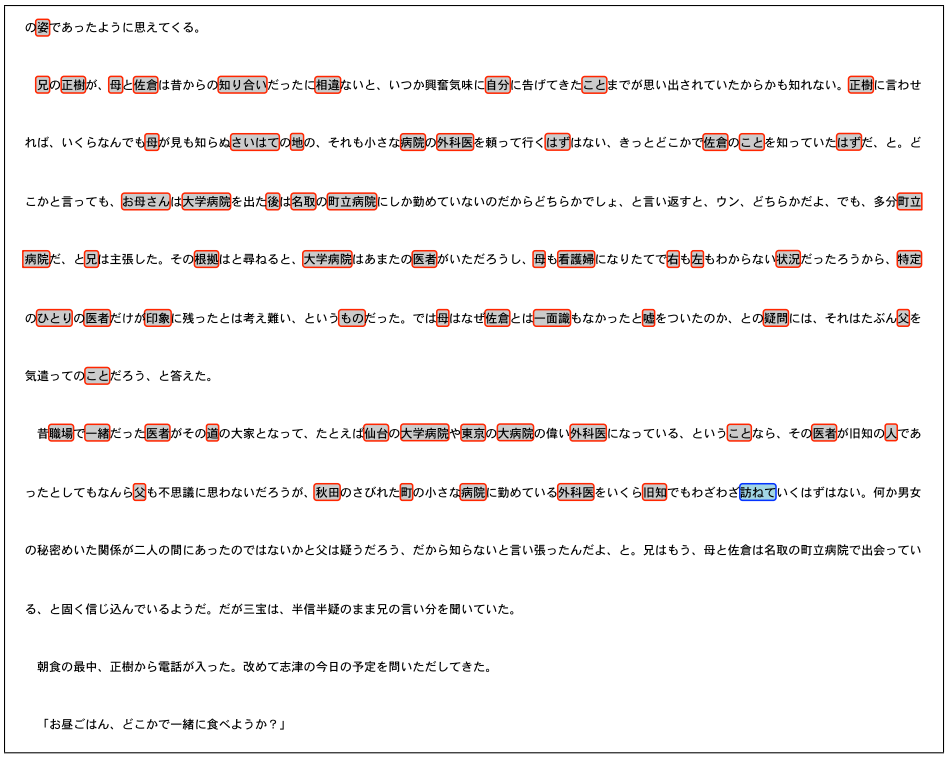

Figure 1: Screenshot of annotation interface

4 Evaluation

4.1 Data

We use the data collected by Mitsuda et al. (2014) for evaluation. Mitsuda et al. (2014) conducted an experiment for collecting annotator's behavioural data during the PAS annotation in Japanese texts. Given a single predicate in a text in which its argument candidates were marked on the screen, the annotators were instructed to identify the *ga* (nominative) argument of the target predicate by clicking it with a mouse. Figure 1 shows a screenshot of the annotation interface, in which a single target predicate in the text is highlighted in blue and *ga* argument candidates are highlighted in gray.

The texts used in their experiment were sampled from Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014). BCCWJ contains approximate 100 million words collected from around 170 thousand texts of various domains. The texts are annotated with various kinds of information at bibliographic and morphological levels. BCCWJ has two types of schema for dividing a sentence into words: Long Unit Word (LUW) and Short Unit Word (SUW). Roughly speaking, an LUW corresponds to a compound noun, whereas an SUW corresponds to a component word of the compound. The LUW were used as argument candidates in their experiment.

BCCWJ consists of three sub-corpora (Publication, Library, and Special-purpose) and each sub-corpus has several registers. For example, PB register, whose texts are sampled from books, is a part of the publication sub-corpus.

About one percent of BCCWJ is defined as the "core data", and the core data is manually annotated with richer linguistic information. The core data annotated with dependency structures, coordinate structures, coreference relations and predicate argument structures is particularly called BCCWJ-DepParaPAS (Ueda et al., 2015; Maekawa et al., 2014). BCCWJ-DepParaPAS adopts the annotation schema of NAIST Text Corpus (Iida et al., 2007b) where three obligatory arguments, *ga* (nominative), *wo* (accusative) and *ni* (dative) are annotated at the SUW level.

The 221 texts used in their experiment were sampled from the PB register of BCCWJ-DepParaPAS, thus they have rich linguistic information including the PAS information. Among these 221 texts we discarded 37 texts because of the discrepancy in the annotation schema between BCCWJ-DepParaPAS and the experiment conducted by Mitsuda et al. (2014). The remaining 184 texts comprise 107 texts with intra-sentence *ga* arguments and 77 texts with inter-sentence ones. The number of candidates including the correct one in a text ranges from 12 to 109, with the median being 60.

In the experiment by Mitsuda et al. (2014), the texts were independently annotated by 20 annotators. During the annotation, annotator eye gaze was captured by the eye tracker Tobii T60 at intervals of 1/60 second. Although recent development of the eye-tracking technology enables us to capture eye gaze handily and precisely, we have still errors in the eye tracking data due to various factors such as the ambient lighting conditions, the annotator's glasses, and the characteristics of annotator's eyes. Among data from 20 annotators, we discarded the data from 13 annotators because their data included a session the tracking error rate of which exceeded 30%. Here "a session" stands for a single annotation session in which an annotator chooses a *ga* argument of the specified predicate in a single text. In summary, we utilise the data consisting of the 184 texts annotated by seven annotators.

Training data We prepare the following three types of models for evaluation and construct the training data for each model from the data explained above.

- **Binary Regression (BiReg) model**:
  This model chooses the argument candidate with the highest regression value as the correct *ga* argument. In the training examples, only the correct argument, i.e. the manually annotated *ga* argument, is considered a positive example, and all other candidates in the text are negative examples.

- **Distance Ranking (DRank) model**:
  This model ranks the argument candidates according to their plausibility of being the *ga* argument of the specified predicate. The distance between the predicate and the candidate is used for calculating the rank. The training examples are created as follows. For each annotation instance, the correct

argument is placed at the top of the ranking, and the other candidates are ranked in the ascending order of their distance feature, i.e. the normalised number of LUWs between the predicate and the candidate. We prepare this model for verifying the effectiveness of the ranking model based on the other metric than the fixation frequency.

- **Fixation Ranking (FixRank) model**:
  This model also ranks the argument candidates according to their plausibility of being the *ga* argument of the specified predicate. Training examples are created based on the number of fixations on the candidates as described in 3.4.

Test data We sampled 29,519 predicate instances from the PB register in the BCCWJ core data: 21,816 intra-sentence instances and 7,703 inter-sentence instances. We kept the range of the number of candidates in a text of the test data as the same as that of the training data, ranging from 12 to 109 with the median being 48.

4.2 Results and discussion

For the training of the BiReg model, we utilised Classias (Okazaki, 2009). As a learning algorithm, L2 regularised logistic regression was adopted. For the training of the DRank and FixRank models, we utilised SVM$^{\text{rank}}$ (Joachims, 2006). As a slack variable, L1-norm was adopted. We prepared four variations of the feature set for model training: a base set (F_{base}), the base set with semantic category features (F_{sem}), the base set with syntactic dependency features (F_{syn}), and the base set with both semantic and syntactic features (F_{synsem}). These feature sets are summarised in Table 2.

Table 2: Variations of feature set

feature set	feature ID defined in Table 1
F_{base}	1–11, 13–16
F_{sem}	1–17
F_{syn}	1–11, 13–16, 18
F_{synsem}	1–18

The all models estimated *ga* likeliness of each candidate, and then selected the candidate with the maximum score as the *ga* argument. If the selected candidate agreed with the answer, it was judged to be correct. Even though the selected candidates were in the same coreference chain as the answer argument, we judged it as wrong selection. We adopted a strict criterion in correctness.

Table 3: Evaluation result (Accuracy)

model	feature set	intra	inter	total
BiReg	F_{base}	0.56	0.04	0.42
	F_{sem}	0.48	**0.06**	0.37
	F_{syn}	0.58	0.03	0.44
	F_{synsem}	0.52	0.05	0.40
DRank	F_{base}	0.47	0.01	0.35
	F_{sem}	0.47	0.01	0.35
	F_{syn}	0.50	0.01	0.37
	F_{synsem}	0.51	0.01	0.38
FixRank	F_{base}	0.55	0.02	0.41
	F_{sem}	0.49	0.02	0.37
	F_{syn}	**0.63**	0.02	**0.47**
	F_{synsem}	0.58	0.02	0.43

Table 3 shows the accuracy of each model and feature set combination. We calculated the accuracy for two groups: intra-sentence cases, i.e. a predicate and its *ga* argument are in the same sentence, and inter-sentence cases, i.e. they are not in the same sentence. Table 3 shows that FixRank+F_{syn} shows the

highest accuracy in the 'intra-sentence' and 'total' columns, while the BiReg model, particularly with F_{sem}, shows the better accuracy in the 'inter-sentence' column.

As far as this result is concerned, the FixRank+F_{syn} model predicts intra *ga* arguments the best of all the models. This supports our hypothesis that frequent looks at argument candidates imply their plausibility of being the argument of the predicate, and utilising eye gaze information contributes to the improvement in parameter estimation of the PAS analysis model.

The table also showed that the semantic feature set F_{sem} does not work well except for the inter-sentence cases with the BiReg model. We automatically assigned the semantic categories to the arguments. As Taira et al. (2008) did, it is worthwhile to see if manual assignment of the semantic categories would improve the effectiveness of F_{sem}.

We need to mention that the DRank model showed the worst performance among three models. The training examples for the DRank model would contain more superfluous information than the FixRank model because the candidates were rigidly ordered in the ascending order of the distance feature, thus there was no candidate of equal rank. This rigid ranking could make the DRank model fail to capture better negative examples.

Table 4: Comparison between BiReg+F_{syn} model and FixRank+F_{syn} model

(15) intra/inter	(18) dependency	(11) case marker	BiReg+F_{syn} P	BiReg+F_{syn} R	BiReg+F_{syn} F	FixRank+F_{syn} P	FixRank+F_{syn} R	FixRank+F_{syn} F
intra	direct	*ga*	0.949	0.884	0.915	0.948	0.975	0.962
		others	0.513	0.463	0.487	0.403	0.693	0.509
	indirect	*ga*	0.227	0.602	0.330	0.242	0.561	0.338
		others	0.184	0.166	0.175	0.156	0.103	0.124
	no	*ga*	0.294	0.892	0.442	0.352	0.741	0.477
		others	0.239	0.469	0.316	0.306	0.290	0.297
inter	no	*ga*	0.232	0.050	0.082	0.284	0.025	0.046
		others	0.247	0.031	0.055	0.172	0.014	0.025

"others" in the case marker column includes arguments without case markers.
P, R, and F denote precision, recall, and F-measure, respectively.

For further analysis, we focused on the specific three features: (11) case marker, (15) intra/inter, and (18) dependency, and calculated precision (P), recall (R), and F-measure (F) for the test cases with each combination of these features. Table 4 shows the comparison between the BiReg+F_{syn} model and the FixRank+F_{syn} model that show the highest accuracy in total in Table 3.

The notable difference between these two model is that the FixRank model shows the better F-measures for directly dependent arguments, particularly in *ga* arguments. Concerning the directly dependent *ga* argument cases, the FixRank shows the better recall value with the similar precision value. In contrast, in the cases of directly dependent 'other' case markers, the BiReg shows the better precision value. We investigated the number of directly dependent arguments that each model selected as the answer. The FixRank model selected 19,612 directly dependent arguments, and 13,751 *ga* argument among them, while the BiReg model selected 12,520 directly dependent arguments with 7,208 *ga* arguments. These numbers suggest that the FixRank model tends to choose directly dependent arguments regardless of their case markers.

Concerning the FixRank model, we further investigated the number of the second rank training candidates that directly depend on the target predicate, i.e. the directly dependent "better negative examples", to obtain 428 candidates out of 1,288 candidates ($= 184$ texts $\times$ 7 annotators) that make 33% of the total training examples. On the other hand, the number of the correct arguments that directly depend on the target predicate is 35 out of 184 examples, making 19% of the texts. These numbers suggest that the annotators tend to look at directly dependent arguments more during the annotation, the FixRank model tends to select directly dependent arguments as our method proposes frequently fixated arguments as "better negative candidates". This tendency would explain the poor performance of the FixRank model in the inter-sentence cases.

5 Conclusion

We proposed utilising the information of annotator eye gaze during their annotation for estimating the parameters of the Japanese PAS analysis model. Through the evaluation in which the *ga* argument of a specified predicate was identified, we confirmed the effectiveness of the eye gaze information for the PAS analysis. We gained 0.05 point increase in accuracy by introducing the eye gaze information into the parameter estimation. The accuracy for the inter-sentence arguments, however, still remains very low. Moreover, the eye gaze information does not work well on the inter-sentence arguments. We need to refine the usage of the eye gaze information for further improvement of the PAS analysis. Currently we use only the fixation frequency, but the fixation duration might provide the information from different aspects. Also instead of utilising all fixations, their selective usage would be another exploring direction.

In the current evaluation, we used the DRank model as the baseline of the ranking model, but its performance is too poor to be a fair baseline. It would be necessary to adopt a state-of-the-art ranking model such as Hangyo et al. (2013) for further evaluation.

The eye gaze information has attracted much attention in various NLP tasks in recent years such as dialogue systems (Prasov et al., 2007; Qu and Chai, 2007), reference resolution (Prasov and Chai, 2008), dependency parsing (Barrett and Søgaard, 2015), coreference resolution (Ross et al., 2016), named entity recognition (Tomanek et al., 2010). Exploring the effectiveness of eye gaze information during annotation in other NLP tasks would be another important research direction.

References

Maria Barrett and Anders Søgaard. 2015. Using reading behavior to predict grammatical functions. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, pages 1–5.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, September.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2013. Japanese zero reference resolution considering exophora and author/reader mentions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 924–934.

Yuta Hayashibe, Mamoru Komachi, and Yuji Matsumoto. 2011. Japanese predicate argument structure analysis exploiting argument position and type. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 201–209.

Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2007a. Zero-anaphora resolution by learning rich syntactic pattern features. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(4):1:1–1:22.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007b. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proceedings of ACL 2007 Workshop on Linguistic Annotation*, pages 132–139.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi. 1997. *Nihongo Goi Taikei, A Japanese Lexicon*. Iwanami Shoten, Tokyo.

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.

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

Thorsten Joachims. 2006. Training Linear SVMs in Linear Time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Marcel Adam Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Mamoru Komachi, Ryu Iida, Kentaro Inui, and Yuji Matsumoto. 2007. Learning-based argument structure analysis of event-nouns in Japanese. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, pages 120–128.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language resources and evaluation*, 48(2):345–371.

Koh Mitsuda, Ryu Iida, and Takenobu Tokunaga. 2013. Detecting missing annotation disagreement using eye gaze information. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 19–26.

Koh Mitsuda, Ryu Iida, and Takenobu Tokunaga. 2014. Collection and analysis of eye gaze information in single predicate-argument relation annotation (in Japanese). In *Proceedings of Special Interest Group on Natural Language Processing (IPSJ-SIGNL)*, volume 2014-NL-217-2, pages 1–9. Information Processing Society of Japan.

Naoaki Okazaki. 2009. Classias: a collection of machine-learning algorithms for classification. `http://www.chokkan.org/software/classias/`.

Hiroki Ouchi, Hiroyuki Shindo, Kevin Duh, and Yuji Matsumoto. 2015. Joint case argument identification for Japanese predicate argument structure analysis. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 961–970.

Zahar Prasov and Joyce Y. Chai. 2008. What's in a gaze?: The role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 20–29.

Zahar Prasov, Joyce Y. Chai, and Hogyeong Jeong. 2007. Eye gaze for attention prediction in multimodal human-machine conversation. In *Proceedings of the AAAI Spring Symposium on Interaction Challenges for Artificial Assistants*, pages 102–110.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly.

Shaolin Qu and Joyce Y. Chai. 2007. An exploration of eye gaze in spoken language processing for multimodal conversational interfaces. In *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 284–291.

Joe Cheri Ross, Abhijit Mishra, and Pushpak Bhattacharyya. 2016. Leveraging annotators' gaze behaviour for coreference resolution. In *Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 22–26.

Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (ETRA 2000)*, pages 71–78.

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.

Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. 2016. Neural network-based model for Japanese predicate argument structure analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1235–1244.

Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. 2008. A Japanese predicate argument structure analysis using decision lists. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 523–532.

Takenobu Tokunaga, Ryu Iida, and Koh Mitsuda. 2013. Annotation for annotation – toward eliciting implicit linguistic knowledge through annotation –. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9)*, pages 79–83.

Katrin Tomanek, Udo Hahn, Steffen Lohmann, and Jürgen Ziegler. 2010. A cognitive cost model of annotations based on eye-tracking data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 1158–1167.

Yoshiko Ueda, Ryu Iida, Masayuki Asahara, Yuji Matsumoto, and Takenobu Tokunaga. 2015. Predicate-argument structure and coreference relation annotation on 'balanced corpus of contemporary written Japanese' (in Japanese). In *Proceedings of the 8th Japanese corpus linguistics workshop*, pages 205–214.

Katsumasa Yoshikawa, Masayuki Asahara, and Yuji Matsumoto. 2011. Jointly extracting Japanese predicate-argument relation with markov logic. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1125–1133.