# Detecting Missing Annotation Disagreement using Eye Gaze Information

**Koh Mitsuda** **Ryu Iida** **Takenobu Tokunaga**
Department of Computer Science, Tokyo Institute of Technology
{mitsudak,ryu-i,take}@cl.cs.titech.ac.jp

## Abstract

This paper discusses the detection of missing annotation disagreements (MADs), in which an annotator misses annotating an annotation instance while her counterpart correctly annotates it. We employ annotator eye gaze as a clue for detecting this type of disagreement together with linguistic information. More precisely, we extract highly frequent gaze patterns from the pre-extracted gaze sequences related to the annotation target, and then use the gaze patterns as features for detecting the MADs. Through the empirical evaluation using the data set collected in our previous study, we investigated the effectiveness of each type of information. The results showed that both eye gaze and linguistic information contributed to improving performance of our MAD detection model compared with the baseline model. Furthermore, our additional investigation revealed that some specific gaze patterns could be a good indicator for detecting the MADs.

## 1 Introduction

Over the last two decades, with the development of supervised machine learning techniques, annotating texts has become an essential task in natural language processing (NLP) (Stede and Huang, 2012). Since the annotation quality directly impacts on performance of ML-based NLP systems, many researchers have been concerned with building high-quality annotated corpora at a lower cost. Several different approaches have been taken for this purpose, such as semi-automating annotation by combining human annotation and existing NLP tools (Marcus et al., 1993; Chou et al., 2006; Rehbein et al., 2012; Voutilainen, 2012), implementing better annotation tools (Kaplan et al., 2012; Lenzi et al., 2012; Marcińczuk et al., 2012).

The assessment of annotation quality is also an important issue in corpus building. The annotation quality is often evaluated with the agreement ratio among annotation results by multiple independent annotators. Various metrics for measuring reliability of annotation have been proposed (Carletta, 1996; Passonneau, 2006; Artstein and Poesio, 2008; Fort et al., 2012), which are based on inter-annotator agreement. Unlike these past studies, we look at annotation processes rather than annotation results, and aim at eliciting useful information for NLP through the analysis of annotation processes. This is in line with *Behaviour mining* (Chen, 2006) instead of data mining. There is few work looking at the annotation process for assessing annotation quality with a few exceptions like Tomanek et al. (2010), which estimated difficulty of annotating named entities by analysing annotator eye gaze during her annotation process. They concluded that the annotation difficulty depended on the semantic and syntactic complexity of the annotation targets, and the estimated difficulty would be useful for selecting training data for active learning techniques.

We also reported an analysis of relations between a necessary time for annotating a single predicate-argument relation in Japanese text and the agreement ratio of the annotation among three annotators (Tokunaga et al., 2013). The annotation time was defined based on annotator actions and eye gaze. The analysis revealed that a longer annotation time suggested difficult annotation. Thus, we could estimate annotation quality based on the eye gaze and actions of a single annotator instead of the annotation results of multiple annotators.

Following up our previous work (Tokunaga et al., 2013), this paper particularly focuses on a certain type of disagreement in which an annotator misses annotating a predicate-argument relation

*International Joint Conference on Natural Language Processing*, pages 19–26,
Nagoya, Japan, 14-18 October 2013.

while her counterpart correctly annotates it. We call this type of disagreement *missing annotation disagreement (MAD)*. MADs were excluded from our previous analysis. Estimating MADs from the behaviour of a single annotator would be useful in a situation where only a single annotator is available. Against this background, we tackle a problem of detecting MADs based on both linguistic information of annotation targets and annotator eye gaze. In our approach, the eye gaze data is transformed into a sequence of fixations, and then fixation patterns suggesting MADs are discovered by using a text mining technique.

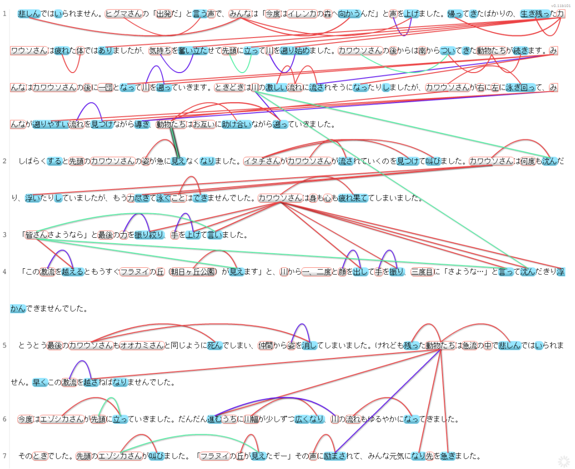
This paper is organised as follows. Section 2 presents details of the experiment for collecting annotator behavioural data during annotation, as well as details on the collected data. Section 3 overviews our problem setting, and then Section 4 explains a model of MAD detection based on eye-tracking data. Section 5 reports the empirical results of MAD detection. Section 6 reviews the related work and Section 7 concludes and discusses future research directions.

## 2 Data collection

### 2.1 Materials and procedure

We conducted an experiment for collecting annotator actions and eye gaze during the annotation of predicate-argument relations in Japanese texts. Given a text in which candidates of predicates and arguments were marked as *segments* (i.e. text spans) in an annotation tool, the annotators were instructed to add links between correct predicate-argument pairs by using the keyboard and mouse. We distinguished three types of links based on the case marker of arguments, i.e. *ga* (nominative), *o* (accusative) and *ni* (dative). For elliptical arguments of a predicate, which are quite common in Japanese texts, their antecedents were linked to the predicate. Since the candidate predicates and arguments were marked based on the automatic output of a parser, some candidates might not have their counterparts.

We employed a multi-purpose annotation tool *Slate* (Kaplan et al., 2012), which enables annotators to establish a link between a predicate segment and its argument segment with simple mouse and keyboard operations. Figure 1 shows a screenshot of the interface provided by *Slate*. Segments for candidate predicates are denoted by light blue rectangles, and segments for candidate arguments are enclosed with red lines. The colour of links corresponds to the type of relations; red, blue and green denote nominative, accusative and dative respectively.



1 悲しんではいられません。ヒグマさんの「出発だ」と言う声で、みんなは「今度はイレンカの森へ向かうんだ」と声を上げました。帰ってきたばかりの、生き残ったカワウソさんは疲れた体ではありましたが、気持ちを奮い立たせて先頭に立って川を遡り始めました。カワウソさんの後からは南からついてきた動物たちが続きます。みんなはカワウソさんの後に一団となって川を遡っていきます。ときどきは川の激しい流れに流されそうになったりしましたが、カワウソさんが右に左に泳ぎ回って、みんなが遡りやすい流れを見つけながら導き、動物たちはお互いに助け合いながら遡っていきました。

2 しばらくすると先頭のカワウソさんの姿が急に見えなくなりました。イタチさんがカワウソさんが流されていくのを見つけて叫びました。カワウソさんは何度も沈んだり、浮いたりしていましたが、もう力尽きて泳ぐことはできませんでした。カワウソさんは身も心も疲れ果ててしまいました。

3 「皆さんさようなら」と最後の力を振り絞り、手を上げて言いました。

4 「この激流を越えるともうすぐフラヌイの丘（朝日ヶ丘公園）が見えます」と、川から一、二度と顔を出して手を振り、三度目に「さような…」と言って沈んだきり浮かんできませんでした。

5 とうとう最後のカワウソさんもオオカミさんと同じように死んでしまい、仲間から姿を消してしまいました。けれども残った動物たちは急流の中で悲しんではいられません。早くこの激流を越さねばなりませんでした。

6 今度はエゾシカさんが先頭に立っていきました。だんだん進むうちに川幅が少しずつ広くなり、川の流れもゆるやかになってきました。

7 そのときでした。先頭のエゾシカさんが叫びました。「フラヌイの丘が見えたぞー」その声に励まされて、みんな元気になり先を急ぎました。

Figure 1: Interface of the annotation tool

| Event label | Description |
|---|---|
| create_link_start | creating a link starts |
| create_link_end | creating a link ends |
| select_link | a link is selected |
| delete_link | a link is deleted |
| select_segment | a segment is selected |
| select_tag | a relation type is selected |
| annotation_start | annotating a text starts |
| annotation_end | annotating a text ends |

Table 1: Recorded annotation events

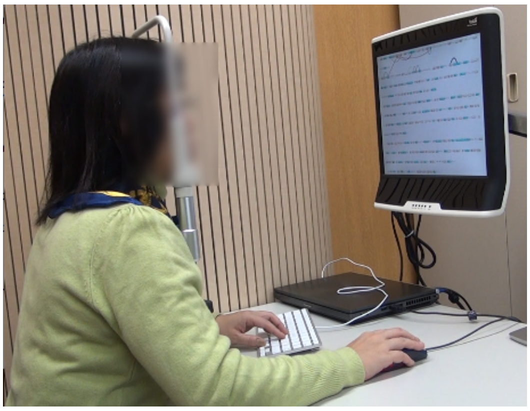


Figure 2: Snapshot of annotation using Tobii T60

In order to collect every annotator operation, we modified *Slate* so that it could record several important annotation events with their time stamp. The recorded events are summarised in Table 1.

Annotator gaze was captured by the Tobii T60 eye tracker at intervals of 1/60 second. The Tobii's display size was 17-inch ($1,280 \times 1,024$ pixels) and the distance between the display and the annotator's eye was maintained at about 50 cm. The five-point calibration was run before starting annotation. In order to minimise the head movement, we used a chin rest as shown in Figure 2.

We recruited three annotators who had experiences in annotating predicate-argument relations. Each annotator was assigned 43 texts for annotation, which were the same across all annotators. These 43 texts were selected from a Japanese balanced corpus, BCCWJ (Maekawa et al., 2010). To eliminate unneeded complexities for capturing eye gaze, texts were truncated to about 1,000 characters so that they fit into the text area of the annotation tool and did not require any scrolling. It took about 20–30 minutes for annotating each text. The annotators were allowed to take a break whenever she/he finished annotating a text. Before restarting annotation, the five-point calibration was run every time. The annotators accomplished all assigned texts after several sessions for three or more days in total.

### 2.2 Results

The number of annotated links between predicates and arguments by three annotators $A_0$, $A_1$ and $A_2$ were 3,353 ($A_0$), 3,764 ($A_1$) and 3,462 ($A_2$) respectively. There were several cases where the annotator added multiple links of the same type to a predicate, e.g. in case of conjunctive arguments; we exclude these instances for simplicity in the analysis below. The number of the remaining links was 3,054 ($A_0$), 3,251 ($A_1$) and 2,996 ($A_2$) respectively. Among them, annotator $A_1$ performed less reliable annotation. Furthermore, annotated *o* (accusative) and *ni* (dative) cases also tend not to be reliable because of the lack of the reliable reference dictionary (e.g. frame dictionary) during annotation. For these reasons, *ga* (nominative) instances annotated by at least one annotator ($A_0$ or $A_2$) are used in the rest of this paper.

## 3 Task setting

Annotating nominative cases might look a trivial task because the *ga*-case is usually obligatory, thus given a target predicate, an annotator could exhaustively search for its nominative argument in an entire text. However, this annotation task becomes problematic due to two types of exceptions. The first exception is exophora, in which an argument does not explicitly appear in a text because of the implicitness of the argument or the referent outside the text. The second exception is functional usage of predicates, i.e. a verb can be used like a functional word. For instance, in the expression "*kare ni kuwae-te* (in addition to him)", the verb "*kuwae-ru* (add)" works like a particle instead of a verb. There is no nominative argument for the verbs of such usage. These two exceptions make annotation difficult as annotators should judge whether a given predicate actually has a nominative argument in a text or not. The annotators actually disagreed even in nominative case annotation in our collected data. The statistics of the disagreement are summarised in Table 2 in which the cell at both "not annotated" denotes the number of predicates that were not annotated by both annotators.

| $A_0 \setminus A_2$ | annotated | not annotated |
|---|---|---|
| annotated | 1,534 | 312 |
| not annotated | 281 | 561 |

Table 2: Result of annotating *ga* (nominative) arguments by $A_0$ and $A_2$

As shown in Table 2, when assuming the annotation by one of the annotators is correct, about 15% of the annotation instances is missing in the annotation by her counterpart. Our task is defined to distinguish these missing instances (312 or 281) from the cases that both annotators did not make any annotation (561).

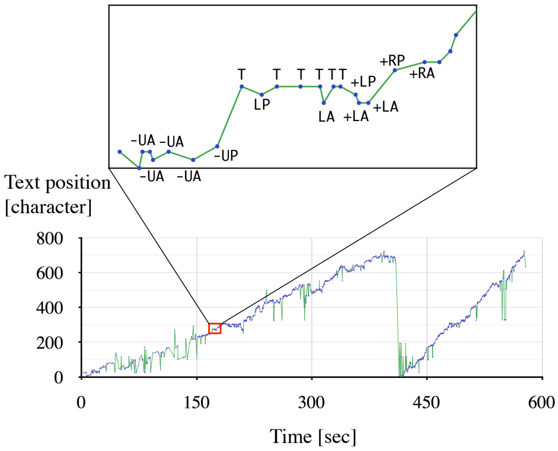


Figure 3: Example of the trajectory of fixations during annotation

## 4 Detecting missing annotation disagreements

We assume that annotator eye movement gives some clues for erroneous annotation. For instance, annotator gaze may wander around a target predicate and its probable argument but does not eventually establish a link between them, or the gaze accidentally skips a target predicate. We expect that some specific patterns of eye movements could be captured for detecting erroneous annotation, in particular for MADs.

To capture specific eye movement patterns during annotation, we first examine a trajectory of fixations during the annotation of a text. The gaze fixations were extracted by using the Dispersion-Threshold Identification (I-DT) algorithm (Salvucci and Goldberg, 2000). The graph in Figure 3 shows the fixation trajectory where the x-axis is a time axis starting from the beginning of annotating a text, and the y-axis denotes a relative position in the text, i.e. the character-based offset from the beginning of the text. Figure 3 shows that the fixation proceeds from the beginning to the end of the text, and returns to the beginning at around 410 sec. A closer look at the trajectory reveals that the fixations on a target predicate are concentrated within a narrow time period. This leads us to the local analysis of eye fixations around a predicate for exploring meaningful gaze patterns. In addition, we focus on the first annotation process, i.e. the time region from 0 to 410 sec in Figure 3 in this study.

Characteristic gaze patterns are extracted from a fixation sequence by following three steps.

1. We first identify a time period for each target predicate where fixations on the predicate are concentrated. We call this period *working period* for the predicate.
2. Then a series of fixations within a working period is transformed into a sequence of symbols, each of which represents characteristics of the corresponding fixation.
3. Finally, we apply a text mining technique to extract frequent symbol patterns among a set of the symbol sequences.

In step 1, for each predicate in a text, a sequence of fixations is scanned along the time axis with a fixed window size. We decided the window size such that the window always covers exactly 40 fixations on any segment. This size was fixed based on our qualitative analysis of the data. The window covering the maximum number of the fixations on the target predicate is determined. A tie breaks by choosing the earlier period. Then the first and the last fixations on the target predicate within the window are determined. Furthermore, we add 5 fixations as a margin before the first fixation and after the last fixation on the target predicate. This procedure defines a working period of a target predicate. Figure 4 illustrates the definition of a working period of a target predicate.

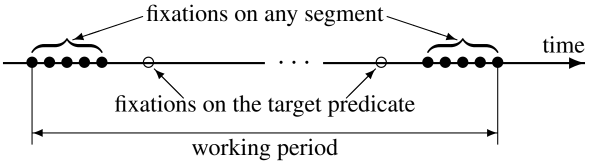


Figure 4: Definition of a working period

| category | symbols |
|---|---|
| position | (U)pper, (B)ottom, (R)ight, (L)eft |
| segment type | (T)arget predicate, other (P)redicate,<br>(A)rgument candidate |
| time period | within the preceding margin (−),<br>within the following margin (+) |

Table 3: Definition of symbols for representing gaze patterns

| | | |
|---|---|---|
| | (U)pper | |
| (L)eft | (T)arget predicate | (R)ight |
| | (B)ottom | |

Figure 5: Definition of gaze areas

In step 2, each fixation in a working period is converted into a combination of pre-defined symbols representing characteristics of the fixation with respect to its relative position to the target predicate, segment type and time point as shown in Table 3. The fixation position is determined according to the areas defined in Figure 5. For instance, a fixation of an argument candidate to the left of the target predicate is denoted by the symbol 'LA'. Accordingly, a sequence of fixations in a working period is transformed into a sequence of symbols, such as '-UA -UA -UA -UA -UP T LP T T T LA T T +LP +LA +LA +RP +RA' as shown in Figure 3.

In step 3, highly frequent patterns of symbols are extracted from the set of symbol sequences

| type | feature | description |
|---|---|---|
| linguistic | is_verb<br>is_adj<br>lemma | 1 if the target predicate is a verb; otherwise 0.<br>1 if the target predicate is a adjective; otherwise 0.<br>lemma of the target predicate. |
| gaze | gaze_pat$_i$ | 1 if gaze pattern$_i$ extracted in Section 4 is contained in a sequence of fixations for the target predicate; otherwise 0. |

Table 4: Feature set for MAD detection

created in step 2 by using the prefixspan algorithm (Pei et al., 2001), which is a sequential mining method that efficiently extracts the complete set of possible patterns. The extracted patterns are used as features in the MAD classification. In addition to the gaze patterns, we introduced linguistic features as well, such as the PoS and lexical information, as shown in Table 4. In particular, lemma of the target predicate is useful for classification because the MAD instances are skewed with respect to certain verbs and adjectives.

## 5 Evaluation

To investigate the effectiveness of gaze patterns introduced in Section 4, we evaluate performance of detecting MADs in our data. In actual annotation review situations for detecting MADs, it is reasonable to assume that an annotator concentrates her/his attention on only non-annotated predicate-argument relations. We therefore conducted a 10-fold cross validation with the data shown in Table 2 except for the instances annotated by both annotators. The evaluation is two-fold, one evaluates the performance of detecting missing annotations of $A_0$, assuming that $A_2$ annotation is the gold standard, i.e. distinguishing 281 positive instances from 561 negative instances, and the other way around.

We used a Support Vector Machine (Vapnik, 1998) with a linear kernel, altering parameters for the cost and slack variables, i.e. `-j` and `-c` options of svm_light [1]. The parameters of the prefixspan algorithm were set so that the maximum size of patterns was 5 and the minimum size of patterns was 3 due to the computing efficiency. We used the top-50 frequent gaze patterns for both positive and negative cases as gaze features.

### 5.1 Baseline model

We employ a simple baseline model, which classifies all instances into the positive, i.e. it should have been annotated with *ga*-case. This corresponds to a typical verification strategy that an annotator checks all instances except for the nominative arguments annotated by herself.

|  | (gold:$A_0$, eval:$A_2$) |  |  | (gold:$A_2$, eval:$A_0$) |  |  |
|---|---|---|---|---|---|---|
|  | R | P | F | R | P | F |
| baseline | 1.000 | 0.358 | 0.527 | 1.000 | 0.333 | 0.500 |
| ling | 0.933 | 0.402 | 0.562 | 0.846 | 0.467 | 0.599 |
| eye | 0.997 | 0.358 | 0.527 | 0.964 | 0.342 | 0.505 |
| ling+eye | 0.750 | 0.404 | 0.525 | 0.829 | 0.403 | 0.542 |

Table 5: Results of detecting MADs

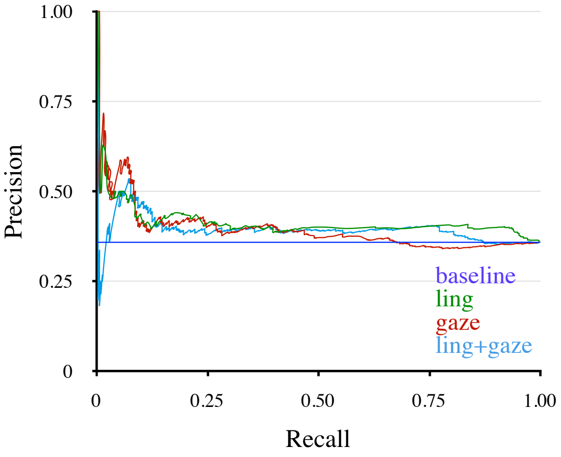


Figure 6: PR-curve (gold:$A_0$, eval:$A_2$)

### 5.2 Results

The results of binary classification are shown in Table 5. The left half shows the evaluation result of $A_2$ with assuming the $A_0$ annotation is the gold standard, and the right half shows the inverse case. The table shows a tendency that any ML-based model outperforms the baseline model, indicating that both linguistic and eye gaze information are useful for detecting MADs. However, combining both information did not work well against our expectation. The results show that the model with only the linguistic features achieved the best performance.

As described in Section 3, we would use the

[1] http://svmlight.joachims.org/

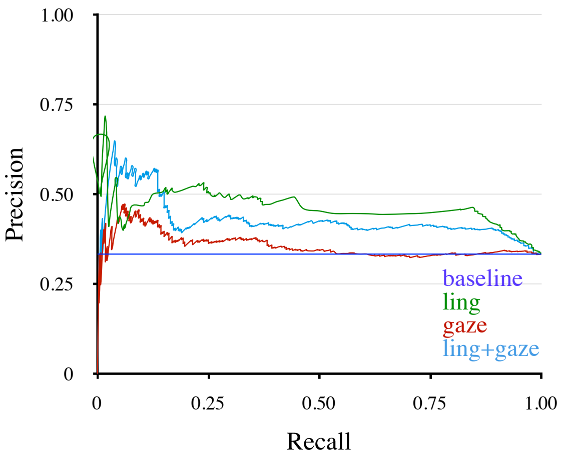


Figure 7: PR-curve (gold:$A_2$, eval:$A_0$)

| freq. | weight | gaze pattern |
|---|---|---|
| 35 | 0.2349 | `T T T` |
| 34 | 0.0258 | `T LA LA` |
| 30 | -0.0510 | `LA LA T` |
| 25 | 0.1220 | `-LP -LP -LP` |
| 25 | 0.0554 | `+RP +RP +RP` |
| 24 | 0.0265 | `-LA -LA T` |
| 22 | 0.1390 | `-LA -LA -LA -LA` |
| 21 | -0.1239 | `LA T T` |
| 20 | 0.0164 | `T T T T` |
| 20 | 0.1381 | `+RA +RA +RA` |
| 18 | 0.0180 | `+RA +RP +RP` |
| 17 | 0.0267 | `-LA -LP -LP` |
| 16 | 0.1023 | `-LA -LA -LA -LA -LA` |
| 14 | 0.1242 | `LA LA LA T` |
| 14 | 0.0045 | `-LP -LP -LA` |
| 13 | 0.1891 | `+RA +RP +RP +RP` |
| 12 | 0.1566 | `RA RP RP` |
| 11 | 0.1543 | `LA LA T T` |
| 10 | 0.0387 | `T LA LA LA` |
| 10 | -0.0629 | `-LA -LA -LA T` |

Table 6: Top-20 frequent gaze patterns (gold:$A_2$, eval:$A_0$)

output of the MAD detection model for revising the annotation results. Thus, ranking instances according to the reliability based on the model outputs is more useful than the categorical classification. From this viewpoint, we re-evaluated the results by inspecting a precision-recall (PR) curve for each model. The PR curves corresponding to Table 5 are illustrated in Figure 6 and Figure 7. The PR curves in Figure 6 are competing, while the curves in Figure 7 show that the model using both linguistic and gaze features achieved better precision at the lower recall area compared with the model using only linguistic features. For further investigation of the results in Figure 7, we examined which gaze patterns were frequently occurred in the instances at the lower recall area. We extracted the instances ranked at lower recall, ranging from 0 to 0.15. Table 6 shows top-20 most frequent gaze patterns with their weight that appeared in these extracted instances. Table 6 reveals several tendencies of human behaviour during annotation. For instance, the pattern ‘`T T T`’ that has the highest positive weight represents that gaze consecutively fixated on the target predicate segment. This could suggest annotator’s deeper consideration on whether to annotate it or not. On the other hand, the patterns ‘`T LA LA`’, ‘`LA LA LA T`’ and ‘`LA LA T T`’, each of which has relatively higher positive weight, correspond to the eye movement which looking back toward the beginning of a sentence for an argument, thus they would frequently happen even though no argument is eventually annotated. This may suggest that an annotator is wondering whether to annotate a probable argument or not.

As seen above, gaze patterns are useful for detecting not all but specific MAD instances. Currently, the parameters and granularity of gaze patterns are heuristically decided based on our intuition and our preliminary investigation. There is still room for improving performance by investigating these issues thoroughly.

## 6 Related work

Recent developments in the eye-tracking technology enables various research fields to employ eye-gaze data (Duchowski, 2002).

Bednarik and Tukiainen (2008) analysed eye-tracking data collected while programmers debug a program. They defined areas of interest (AOI) based on the sections of the integrated development environment (IDE): the source code area, the visualised class relation area and the program output area. They compared the gaze transitions among these AOIs between expert and novice programmers. Since the granularity of their AOIs is coarse, it could be used for evaluate programmer’s expertise, but hardly explain why the expert transition pattern realises a good programming skill. In order to find useful information for language processing, we employed smaller AOIs at the character level.

Rosengrant (2010) proposed an analysis method named *gaze scribing* where eye-tracking data is combined with subjects thought process derived by the think-aloud protocol (TAP) (Ericsson and Simon, 1984). As a case study, he analysed a process of solving electrical circuit problems on the computer display to find differences of problem solving strategy between novice and expert subjects. The AOIs are defined both at a macro level, i.e. the circuit, the work space for calculation, and at a micro level, i.e. electrical components of the circuit. Rosengrant underlined the importance of applying gaze scribing to the solving process of other problems. Although information obtained from TAP is useful, it increases her/his cognitive load, thus might interfere with her/his achieving the original goal.

Tomanek et al. (2010) utilised eye-tracking data to evaluate a degree of difficulty in annotating named entities. They are motivated by selecting appropriate training instances for active learning techniques. They conducted experiments in various settings by controlling characteristics of target named entities. Comparing to their named entity annotation task, our annotation task, annotating predicate-argument relations, is more complex. In addition, our experimental setting is more natural, meaning that all possible relations in a text were annotated in a single session, while each session targeted a single named entity (NE) in a limited context in the setting of Tomanek et al. (2010). Finally, our fixation target is more precise, i.e. words, rather than a coarse area around the target NE.

## 7 Conclusion

This paper discussed the task of detecting the missing annotation disagreements (MADs), in which an annotator misses annotating an annotation target. For this purpose, we employed eye gaze information as well as linguistic information as features for a ML-based approach. Gaze features were extracted by applying a text mining algorithm to a series of gaze fixations on text segments. In the empirical evaluation using the data set collected in our previous study, we investigated the effectiveness of each type of information. The results showed that both eye gaze and linguistic information contributed to improving performance of MAD detection compared with the baseline model. Our additional investigation revealed that some specific gaze patterns could be a good indicator for detecting the disagreement.

In this work, we adopted an intuitive but heuristic representation for fixation sequences, which utilised spatial and temporal aspects of fixations as shown in Table 3 and Figure 5. However, there could be other representation achieving better performance for detecting erroneous annotation. Our next challenge as future work is to explore better representations of gaze patterns for improving performance.

## References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Roman Bednarik and Markku Tukiainen. 2008. Temporal eye-tracking data: Evolution of debugging strategies with multiple representations. In *Proceedings of the 2008 symposium on Eye tracking research & applications (ETRA '08)*, pages 99–102.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Zhengxin Chen. 2006. From data mining to behavior mining. *International Journal of Information Technology & Decision Making*, 5(4):703–711.

Wen-Chi Chou, Richard Tzong-Han Tsai, Ying-Shan Su, Wei Ku, Ting-Yi Sung, and Wen-Lian Hsu. 2006. A semi-automatic method for annotating a biomedical proposition bank. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora*, pages 5–12.

Andrew T. Duchowski. 2002. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, and Computers*, 34(4):455–470.

K. Anders Ericsson and Herbert A. Simon. 1984. *Protocol Analysis – Verbal Reports as Data –*. The MIT Press.

Karën Fort, Claire François, Olivier Galibert, and Maha Ghribi. 2012. Analyzing the impact of prevalence on the evaluation of a manual annotation campaign. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1474–1480.

Dain Kaplan, Ryu Iida, Kikuko Nishina, and Takenobu Tokunaga. 2012. Slate – a tool for creating and maintaining annotated corpora. *Journal for Language Technology and Computational Linguistics*, 26(2):89–101.

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT annotation tool. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 333–338.

Kikuo Maekawa, Makoto Yamazaki, Takehiko Maruyama, Masaya Yamaguchi, Hideki Ogura, Wakako Kashino, Toshinobu Ogiso, Hanae Koiso, and Yasuharu Den. 2010. Design, compilation, and preliminary analyses of balanced corpus of contemporary written Japanese. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2010)*, pages 1483–1486.

Michał Marcińczuk, Jan Kocoń, and Bartosz Broda. 2012. Inforex – a web-based tool for text corpus management and semantic annotation. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 224–230.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*, pages 831–836.

J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M-C. Hsu. 2001. PrefixSpan: Mining sequential patterns efficiently by prefixprojected pattern growth. In *Proceedings 2001 International Conference Data Engineering (ICDE'01)*, pages 215–224.

Ines Rehbein, Josef Ruppenhofer, and Caroline Sporleder. 2012. Is it worth the effort? assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. *Language Resources and Evaluation*, 46(1):1–23.

David Rosengrant. 2010. Gaze scribing in physics problem solving. In *Proceedings of the 2010 symposium on Eye tracking research & applications (ETRA '10)*, pages 45–48.

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 '00)*, pages 71–78.

Manfred Stede and Chu-Ren Huang. 2012. Interoperability and reusability: the science of annotation. *Language Resources and Evaluation*, 46(1):91–94.

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.

V. N. Vapnik. 1998. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing Communications, and control. John Wiley & Sons.

Atro Voutilainen. 2012. Improving corpus annotation productivity: a method and experiment with interactive tagging. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2097–2102.