IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Your Cursor Reveals: On Analyzing Workers' Browsing Behavior and Annotation Quality In Crowdsourcing Tasks

PEI-CHI LO[1] **AND EE-PENG LIM.**[2]
[1,2]School of Computing and Information Systems, Singapore Managment University 178902 Singapore
[1](e-mail: pclo.2017@phdcs.smu.edu.sg)
[2](e-mail: eplim@smu.edu.sg)

Corresponding author: Pei-Chi Lo (e-mail: pclo.2017@phdcs.smu.edu.sg).

ABSTRACT In this work, we investigate the connection between browsing behavior and task quality of crowdsourcing workers performing annotation tasks that require information judgements. Such information judgements are often required to derive ground truth answers to information retrieval queries. We explore the use of workers' browsing behavior to directly determine their annotation result quality. We hypothesize *user attention* to be the main factor contributing to a worker's annotation quality. To predict annotation quality at the task level, we model two aspects of task-specific user attention, also known as *general* and *semantic user attentions*. Both aspects of user attention can be modeled using different types of browsing behavior features but most previous research mostly focuses on the former. This work therefore proposes to model semantic user attention by capturing the worker's understanding of task content using task-semantics specific behavior features. We develop a web-based annotation interface for gathering user behavior data when workers perform a knowledge path retrieval task. With the collected data, we train several prediction models using behavior features corresponding to different aspects of user attention and conduct experiments on a set of annotation tasks performed by 51 Amazon Mechanical Turk workers. We show that the prediction model using both general and semantic user attention features can achieve the best performance of nearly 75% accuracy.

INDEX TERMS Crowdsourcing, Machine Learning, Annotations, User Modeling, Empirical Study

I. INTRODUCTION

MANY computer science researchers have hired crowdsourcing workers to contribute large amount of annotation data for various information tasks in recent years, and this trend is expected to increase. Annotation through crowdsourcing works well when the hired workers produce high quality annotation results. Studies have shown that non-expert annotations at crowdsourcing platforms such as Amazon Mechanical Turk (AMT) can produce quality comparable to that of experts [1]. Nevertheless, different workers produce annotations of varying qualities. Hence, much research has been conducted on modeling and evaluating workers' reliability.

Among the research works on modeling the reliability of annotation workers, some methods require the self-reported expertise and/or psychological/demographic attributes of the workers [2]–[4] which are relatively difficult to obtain. Others determine the workers' expertise through analysing their annotation histories [5]–[7]. All these methods, nevertheless, have overlooked the *cold-start user* and *task-specific annotation quality* issues. The former refers to workers with little or no history data known to the AMT task requester[1]. Cold-start users or workers are very common, as annotation tasks are often assigned to workers new to the task requester. The lack of history data about workers prevent the latter from learning their annotation quality accurately using the existing methods [8]–[11]. Meanwhile, annotation quality of the same worker can also differ substantially when given

[1] The AMT platform may however have complete annotation history data about the workers.

different tasks. Other than tasks having different difficulty levels, there are many other factors that can affect a worker's performance. To our knowledge, most task-specific worker's reliability research works can only model a small subset of these factors. These works again are not cold-start user issue Instead of determining all factors affecting , we are not able to effectively derive task-specific annotation results from different workers. For example, when worker A is more reliable then worker B on the common task t, it is reasonable to give A's answer a higher weight than B's answer before combining the answers.

This paper therefore aims to address both cold-start users and task-specific quality using *online browsing behavior data* of AMT workers which is easy to collect, even for workers without any annotation history. Browsing behavior data, specifically cursor movement data, has been widely studied in Human Computer Interaction (HCI) and user interface (UI) design domains to improve user interface design. Compared with eye-tracking trajectory data, cursor trajectory may be less accurate but can be easily recorded. Other than analysing browsing behavior data to determine whether a user pays attention to on the user interface screen [12], [13], there are works that infer a worker's overall quality (or reliability) based on behavioral data. To our knowledge, all these works focus on a worker's overall quality rather than task-specific quality [14]. Given that even a reliable worker may perform poorly on difficult tasks, it is still important to study task-specific quality based on browsing behavior captured. By connecting task quality of a worker with his/her browsing behavior, we hope to give requesters a new and unique indicator of the worker's task-specific quality for improving their task design, task assignment and worker engagement.

Thus, we aim to answer three research questions related to a worker's task-specific quality through his/her observed browsing behavior. We first ask is it possible for a worker's task-specific quality be inferred based on the worker's behavior that is related to his/her general attention. Features extracted from the browsing behavioral data to capture general user attention have been studied in previous works but such user attention modeling has been studied in non-crowdsourcing settings. We next ask if the worker's attention to the content semantics of a crowdsourcing task could improve the worker's annotation quality. To answer the question, we need to introduce semantic user attention and define for it a set of new features related to browsing of task semantics within the annotation task user interface. As there are no publicly available datasets for task-specific annotation quality, we need to develop a set of annotation tasks and collect annotation behavior data from AMT workers on a specially instrumented crowdsourcing platform. Finally, we ask if it is possible for the prediction of task-specific annotation quality at the early stage of annotating a task with reasonable accuracy.

As we answer the above research questions, we have made the following contributions: (1) the establishment of the relationship between workers' browsing behavior and their task-specific annotation qualities, (2) the development of predictive models which include specially designed semantic attention features to determine workers' quality for specific tasks using their browsing behavioral data in an information judgement task, (3) the experimental findings that show semantic attention features when combined with other general attention related features can predict annotation quality with higher accuracy, and (4) the study of how worker's task-specific quality can be predicted early enough to trigger worker's assistance or intervention.

II. RELATED WORK

Users' browsing behavior has been studied for multiple aims. In this section, we review research that analyses user browsing behavior data for *user attention modeling*, *reading content modeling*, and *user profile modeling*. In addition, we survey how previous works determine task-specific annotation quality with behavior data.

A. USER ATTENTION MODELING

While early studies focus on utilizing eye-trackers to study user attention when browsing parts of websites [15]–[17], recent studies often aim at modeling using mouse cursor behavior in various applications due to the cheaper cost and easier collection of data [13], [18]–[25]. Guo et al. proposed to utilize aggregated and descriptive mouse cursor data such as scrolling and hovering to infer users' query intent [26], [27]. Arapakis et al. utilized descriptive features, aggregated features, and component interaction features of cursor movement to analyse users' within-content engagement such as predicting user interest [28]. In their follow-up works, newly proposed task-specific component interaction features and Recurrent Neural Network-based cursor trajectory representation are proposed for the modeling of user attention [28]–[30].

Search engine is often the application context when past researchers studied user attention modeling [31], [32]. User attention modeling for crowdsourcing applications, in contrast, has not been studied. Unlike search applications, each crowdsourcing worker is expected to understand the task, to demonstrate knowledge relevant to the task, and to make good judgement when the worker performs an annotation task correctly. In our research, we believe a worker's attention plays an important part in performing well all the above activities. We therefore would like to explore user attention related cursor movement features to predict annotation quality.

B. READING CONTENT MODELING

There have been works suggesting the correspondence between cursor moving pattern and user reading behavior, which is relevant to our annotation task which involves a worker understanding an input article content in order to give the correct annotation label. Research on analysing cursor behavior when reading content is thus relevant to our work [12], [33], [34]. One work that is similar to our information judgement setting is by Hauger et al. [35]. The work designed a set of questions related to an article about the game "Go". Users are then required to answer these questions using the article. From this study, it was found that it possible to predict whether the user has read certain parts of the article using client-side interaction such as cursor movement and clicks. The above works however did not look into the prediction of question-specific answer quality provided by a worker.

C. USER PROFILE MODELING USING BROWSING BEHAVIOR DATA

Inferring user gender, personality characteristics, and even emotion status using browsing behavior data have been a popular research topic [36], [37]. Liu et al. conducted a survey to collect feedback from users on how satisfied they are on the search results, and proposed to predict their satisfactory level using mouse movement data [38]. Yamauchi and Xiao proposed to learn a user's emotion status by analysing how far away his/her cursor trajectory is from the shortest path between the cursor's initial position to the submit button [39]. In the crowdsourcing context, Bron et al. studied how one can accurately predict if a user is a fast or slow worker, and infer his/her personality traits using cursor clicking and movement data in a visual search task [40]. Fu et al. and Kwok et al. focused on predicting a crowdsourcing worker's next action given his/her past cursor movement [41], [42]. Both the above studies however did not link worker's efficiency and next action to his/her annotation result quality. Mok et al. proposed to model worker's reliability using cursor trajectory data [14]. While worker's overall reliability could affect task-specific quality, it still does not predict task-specific quality as the worker performs annotation.

D. DETERMINATION OF TASK-SPECIFIC ANNOTATION QUALITY

Rzeszotarski and Kittur found that task fingerprints of crowdsourcing workers, which include the total number of mouse clicks and cursor moves made by the workers, can be used to predict their annotation task quality [43]. Han et al. proposed Wernicke, a crowdsourcing system that also supports worker's task-specific quality assessment using browsing behavior data [44]. They proposed 8 new features on top of the 6 derived from [43], and categorized them into four types, namely temporal behavior features, page navigation behavior features, context behavior features, and compound behavior features. While these works broke new ground to perform task-specific quality assessment with browsing behavior data, they did not include detailed analysis on why the features are helpful. They also focused on very basic descriptive features of the cursor movement data without considering the different segments of the task interface. In this work, we approach the worker quality assessment using behavior-based user attention models. We provide justification for the connection between workers' quality and their behavioral data. Our subsequent studies also show that attention-derived models outperform the baseline models that consider only descriptive features of the cursor.

III. RESEARCH OBJECTIVES

This research seeks to model task-specific annotation quality based on *general user attention* and *semantic user attention* of workers captured in their browsing behavior. General user attention covers the overall attention a user gives to an user interface without considering the task-specific information. Studies have shown that general user attention improves human performance in different work and play activities [45], [46]. We hypothesize that crowdsourcing is another work-related activity that demands user attention for carrying out the tasks well.

A. MODELING OF SEMANTIC USER ATTENTION

Semantic user attention, in contrast, refers to the attention user pays to task semantics embedded in an annotation task. Ignoring or misunderstanding such task semantics will likely lead to wrong annotation results. Consider the annotation task example: "What is the publication year of Crazy Rich Asians". If a worker does not pay attention to the semantics of the task or misunderstands the question, he or she may return the year the "Crazy Rich Asians" movie was released instead of the year the "Crazy Rich Asians" book was published. Task semantics can be different for different information retrieval annotation tasks. For web search queries, both the query and result page content form the semantics of the task. For question answering, the question content and candidate answer constitute the task semantics. We thus expect a worker to perform careful reading of the given query (or question) content and the candidate result (or answer) content, and to match them as part of his/her semantic user attention.

While identifying the browsing behavior features to measure semantic user attention of crowdsourcing workers is important, this research has yet to be studied in the literature. Our research goal is therefore to identify a set of generic browsing behavior features to represent semantic user attention and to study how well semantic user attention can be used to predict a worker's performance on an annotation task.

B. RESEARCH QUESTIONS

With both general and semantic user attention measured by browsing behavior features, we can move on to answer the following key research questions.

- **Research Question 1:** Are we able to then establish the association between the workers' behavior and their annotation task quality based on general user attention?
- **Research Question 2:** Suppose we represent semantic user attention in a task by matching the semantics of query with the semantics of candidate result semantics attended by the worker. Do these semantic attention features contribute to determining the worker's annotation quality?

- **Research Question 3:** The ability to determine a worker's annotation to be correct (or incorrect) from his/her browsing behavior in the middle of task annotation can bring about new improvements to annotation task design, task assignment, and worker engagement. We thus want to ask: "Can we predict the quality of an annotation task well before it is completed? How soon can we predict the annotation task quality with reasonable accuracy?"

To answer Research Question 1, one has to determine browsing behavior features that effectively capture general worker attention. As there are very few works on measuring general user attention with browsing behavior features, we turn to works that study browsing behavior features for modeling a user's emotion status [47], level of attentiveness [48], [49], and confidence [50], [51]. Research Question 2 focuses on the worker's attention to task semantics during annotation. The challenge here is to model semantic user attention by features of browsing behavior data capturing user attention on matching query and candidate result as the worker performs a annotation task, and evaluating the prediction models using semantic user attention features.

Finally, instead of analyzing task quality at the end of worker's annotation, one can start evaluating the worker's annotation quality in the middle of annotation. This opens up the possibility to intervene the annotation task before its completion, e.g., aborting the task, giving extra help to the workers, etc.. While this sounds good, it is unclear when is the right moment to begin the analysis. We therefore aim to design an experiment to answer Research Question 3.

While the general attention and semantic attention may vary for different annotation tasks, we believe there are still generic browsing behavior features that can be used to measure them. To conduct this research, we select a target information judgement task. Based on the target information retrieval task, we design and implement an annotation user interface for workers to annotate candidate results of a set of query tasks in a crowdsourcing study. The interface is also equipped with browsing behavior tracking capabilities to gather the features for measuring general and semantic attention.

IV. CROWDSOURCING STUDY

In this section, we define a target information retrieval problem called *Contextual Path Retrieval*. To obtain the ground truth results of CPR queries using crowdsourcing, we have designed an annotation user interface. Our crowdsourcing study was conducted on Amazon Mechanical Turk (AMT) using our custom-built web-based interface. All of the human subjects involved are informed about the experiments, and have provided consent for us to use their data in the study.

A. CONTEXTUAL PATH RETRIEVAL TASK

Contextual path retrieval (CPR) aims to return the path involving entities and relations of a knowledge graph for explaining the semantic connection between two query entities mentioned in an input text. CPR can be used in applications which require direct or indirect connections between entities mentioned in an article (e.g., news and comprehension passage). We assume that a knowledge graph covering entities and relations between the entities has been given. As there may be multiple paths in the knowledge graph connecting the two query entities, workers are required to determine the ground truth path using a custom built web-based annotation interface shown in Figure 1. The figure shows a news article in the *article section* where the mentions of the two query entities (i.e., "Alfonso Cuarón" and "Children of Men") are highlighted in orange. The article is segmented into sentences for better tracking of the worker's browsing behavior. A candidate path (i.e., Alfonso Cuarón $\xrightarrow{director}$ Children of Men) is shown in the *task section* and the worker is required to judge whether it correctly describes the semantics connecting the two entities. The *Wikipedia iframe section* displays the Wikipedia page describing any Wikipedia-linked entity in the article and task sections that are selected by the worker. If the workers are not familiar with the task-related entities, he/she can click on the entity mention in the article (e.g., "Children of Men") to show its Wikipedia page in the iframe.

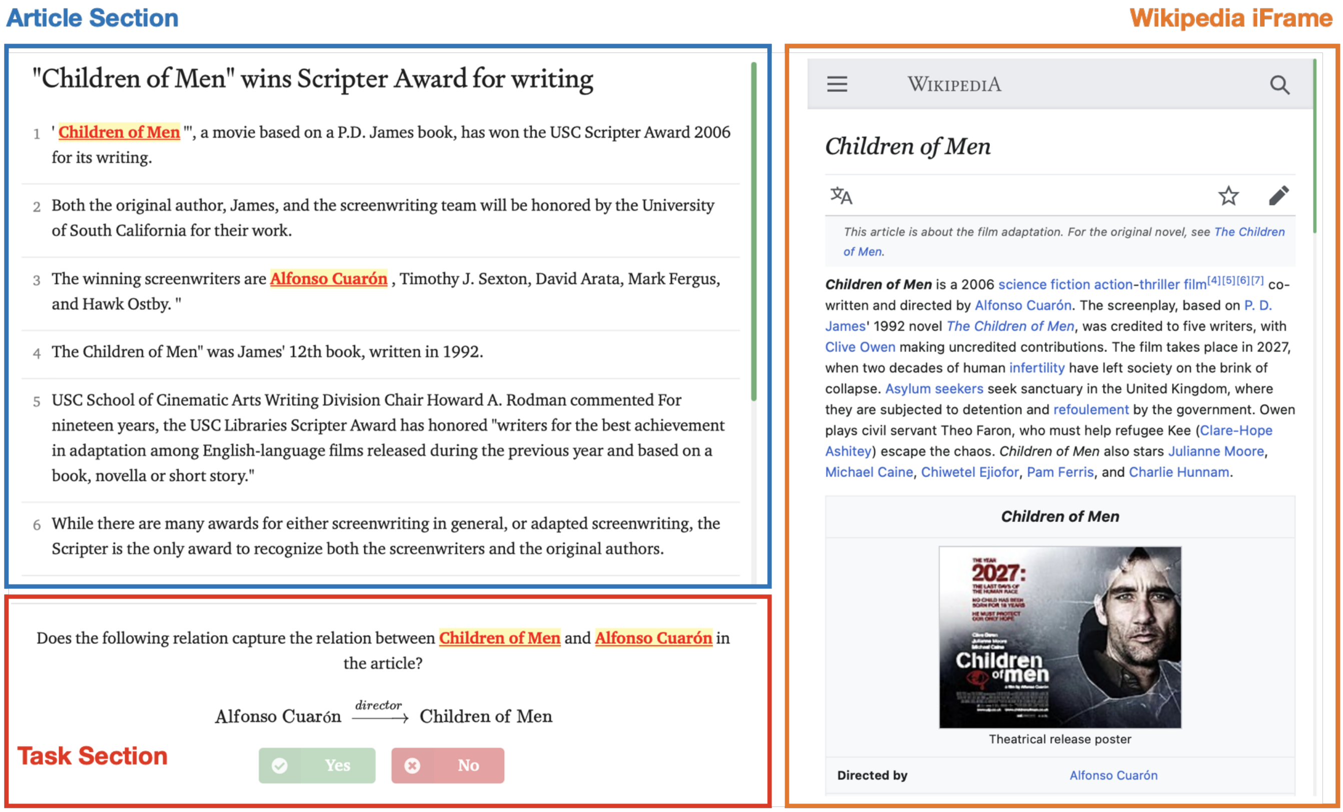
During the annotation of a task, we track the worker's mouse interaction with the elements in the user interface, including the coordinates of the cursor, clicking, hovering, and highlighting events. The interval of such events are also recorded. We also determine the ground truth contextual paths for several query entity pairs found in a set of news articles. With both collected workers' browsing behavior and ground truth data, we carry out this research on the prediction of worker's task-specific annotation quality.

In this study, we use DBpedia[2] as our knowledge graph. DBpedia is a knowledge graph which derives entities and relations from Wikipedia[3]. It is easy to find DBpedia entities mentioned in Wikipedia articles. We design the annotation tasks using a set of 10 Wikinews[4] articles. Wikinews is ideal for a number of reasons: (1) Wikinews articles are well written, (2) they are already classified by topic, (3) they are Wikified, that is, the entity mentions are linked to the Wikipedia entries, and (4) the entities and relations in Wikinews can be found in DBpedia. In this study, we only focus on articles under the Film category. The selected 10 articles are of similar length and require roughly the same amount of time to read through.

From the 10 selected Wikinews articles, we extract 39 query entity pairs and their candidate knowledge paths from the knowledge graph. Each pair of query entities is associated with a Wikinews article mentioning the entities as well as one of the candidate paths which together form an annotation task with the "yes" and "no" answer options. The correct knowledge path of each entity pair is manually determined by an oracle (an author of this paper who is familiar with

[2] https://wiki.dbpedia.org/about
[3] https://en.wikipedia.org/wiki/Main_Page
[4] https://en.wikinews.org/wiki/Main_Page

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3212080



"Children of Men" wins Scripter Award for writing

1 ' Children of Men '", a movie based on a P.D. James book, has won the USC Scripter Award 2006 for its writing.

2 Both the original author, James, and the screenwriting team will be honored by the University of South California for their work.

3 The winning screenwriters are Alfonso Cuarón , Timothy J. Sexton, David Arata, Mark Fergus, and Hawk Ostby. "

4 The Children of Men" was James' 12th book, written in 1992.

5 USC School of Cinematic Arts Writing Division Chair Howard A. Rodman commented For nineteen years, the USC Libraries Scripter Award has honored "writers for the best achievement in adaptation among English-language films released during the previous year and based on a book, novella or short story."

6 While there are many awards for either screenwriting in general, or adapted screenwriting, the Scripter is the only award to recognize both the screenwriters and the original authors.

Does the following relation capture the relation between Children of Men and Alfonso Cuarón in the article?

Alfonso Cuarón $\xrightarrow{director}$ Children of Men

Yes No

WIKIPEDIA

Children of Men

This article is about the film adaptation. For the original novel, see The Children of Men.

Children of Men is a 2006 science fiction action-thriller film[4][5][6][7] co-written and directed by Alfonso Cuarón. The screenplay, based on P. D. James' 1992 novel *The Children of Men*, was credited to five writers, with Clive Owen making uncredited contributions. The film takes place in 2027, when two decades of human infertility have left society on the brink of collapse. Asylum seekers seek sanctuary in the United Kingdom, where they are subjected to detention and refoulement by the government. Owen plays civil servant Theo Faron, who must help refugee Kee (Clare-Hope Ashitey) escape the chaos. *Children of Men* also stars Julianne Moore, Michael Caine, Chiwetel Ejiofor, Pam Ferris, and Charlie Hunnam.

Children of Men

Theatrical release poster

Directed by	Alfonso Cuarón

FIGURE 1. The Annotation Interface

TABLE 1. Statistics of Workers' Behavior from Our Crowdsourced Dataset

Items	Mean	Std
Time Spent Completing A Task	12.3	4.84
Time Spent Completing A Fresh Task	21.3	3.15
Time Spent Completing A Continued Task	10.1	1.27
% Word Hovered	73.2	12.68
Number of Clicks on Interface	8.11	1.33
Number of Highlights	11.1	2.72
Number of Browsed Wikipedia Pages	2.3	0.68

film topic and article content).

To ensure that the workers have read the article thoroughly before performing annotation, they were not allowed to submit answers until they scroll through all sentences of the article. In our task assignment, the same worker might be asked to determine the knowledge paths of different pairs of entities mentioned in the same Wikinews article in different annotation tasks. When this happens, we call the first task involving a fresh article and the subsequent task involving a previously read article the *fresh task* and *continued task* respectively.

B. BEHAVIORAL DATA COLLECTION

To gather browsing behavior data during annotation, we instrument our annotation user interface with cursor tracking capabilities. We use JavaScript and jQuery to capture the hovering, clicking and highlighting events in the article and task sections. Whenever any above-mentioned event occurs, we log the following: element of interaction (e.g., DOM element attributes, xpath of the DOM element), timestamp, and event name.

In addition, we recorded the cursor position every 50 milliseconds during our study. Each position consists of the x and y-axis values. Although we have asked the AMT workers to perform the annotation tasks using desktop PC or laptop, the screen sizes of these devices might still be different. We therefore derived the relative positions of cursor movement. We also recorded the size of the web page when the annotation interface was loaded, and when the web page window was resized[5]. All these browsing behavior data collected were stored in JSON format.

We recruited several AMT workers aged 18 to 43. These workers were all proficient in English (75.3% being native speakers), and were required to at least attain a education background higher than college diploma/ university degree. In total, each worker received 10 tasks to complete in one session. No worker is allowed to join this study multiple times. On average, our crowdsourced workers completed each task in 12.3 seconds (with standard deviation $\sigma = 4.84$). Fresh tasks take longer average time (with mean $\mu = 21.3$,

[5] Nevertheless, we did not record any resizing events throughout our study

standard deviation $\sigma = 3.15$) while continued tasks require shorter time ($\mu = 10.1, \sigma = 1.27$). During annotation, the worker on average hovered over 73.2% of the words in the article, clicked on the interface for 8.11 times, and highlight some parts of interface for 11.1 times. We show statistic in detail in Table 1. This suggests that the workers are generally well engaged during the annotation tasks.

In total, we constructed 39 annotation tasks involving 10 Wikinews articles. 49 AMT workers were recruited. Each worker is assigned 10 of the 39 annotation tasks involving at most 3 articles which are selected randomly. We finally obtained 490 annotations from the workers. When checked against the ground truth paths, we have determined 274 of these annotations to be correct[6] and 216 to be incorrect. That is, the workers achieved 55.9% accuracy.

V. SEMANTIC ATTENTION: AN EMPIRICAL ANALYSIS

A. ILLUSTRATIVE EXAMPLE TASK

In this section, we explore how the workers' behavior affects their annotation quality. We show the heatmaps representing the average time two groups of workers spent hovering/clicking objects in an example task[7] in Figure 2. Worker group A are those who annotated this task correctly ($N = 24$) while group B provided incorrect annotations ($N = 19$). The head and tail entities (i.e., *He's just not that into you* and *Adam Shankman*) are in yellow color. The ground truth contextual path also involves entities *Drew Barrymore* and *Going the Distance*. The green highlights indicates the entity mentions workers paid attention to. The darker the green color is, the more interactions the workers have with the entity mentions.

B. SEMANTIC ATTENTION HEATMAP

Figure 2 essentially shows that the semantic attention heatmaps of groups A and B workers are quite disparate. While the head and tail entity mentions receive the most attention by both worker groups[8], it is not the case for other entity mentions.

The heatmaps show that the workers spend more times interacting with entity mentions that they think are more important and relevant to the task. The question then becomes, can we predict the correctness of an annotation based on the worker's interaction with certain elements?

C. HYPOTHESIS TESTING

We start from identifying entity mentions that affect the workers annotation decision. We hypothesize that workers who spend more time on entity mentions that are more related to the task may more likely annotate correctly. These include the known head and tail entity mentions, and mentions of entities in the ground truth contextual path which are unknown to the workers.

We first conduct a t-test to compare the *time spent hovering over head and tail entity mentions* between group A and group B workers. The two-tailed p-value is 0.3285 (t=0.9869), which suggest there is no statistically significant difference between the two groups of workers. This result shows that semantic attention on head and tail entity mentions is not significantly different between the two worker groups.

Next, we conduct another similar test on *time spent hovering over all mentions of all ground truth path entities (head and tail included)* between groups A and B users. These entities might be implicit to the workers as some tasks show only the non-ground truth contextual paths for worker annotations. Our t-test yields a two-tailed p-value of 0.0001 (t=4.5070). The null hypothesis is rejected and we conclude there exist statistically significant difference between the two groups of worker in their semantic attention on entities of the ground truth path. Group A workers have more interaction with mentions of such entities, which may result in better annotation accuracy.

Finally, we want to determine if workers spending time on mentions of entities that are neither head/tail entity nor contextual path entities could contribute to annotation accuracy. Our t-test result with p-value 0.0554 (t=1.9715) however suggests no significant difference between groups A and B workers for this behavior.

In conclusion, worker's interaction with contextual path entities is the only behavioral feature that shows significant difference between workers of high and low correctness. This attention feature captures the worker's ability to identify important entity mentions and understand the actual semantic relationship between the task entities(i.e., head and tail entities). While we find significant correlation between the interaction with contextual path entities and correctness, the ground truth path is not given under real-world setting. As a result, we should design features that represent the unobserved ground truth contextual path entities with observed entities. We will elaborate how we design the semantic attention features based on these findings in Section VI.

VI. PREDICTION OF ANNOTATION QUALITY USING BROWSING BEHAVIOR FEATURES

A. BEHAVIOR FEATURES FOR PREDICTION

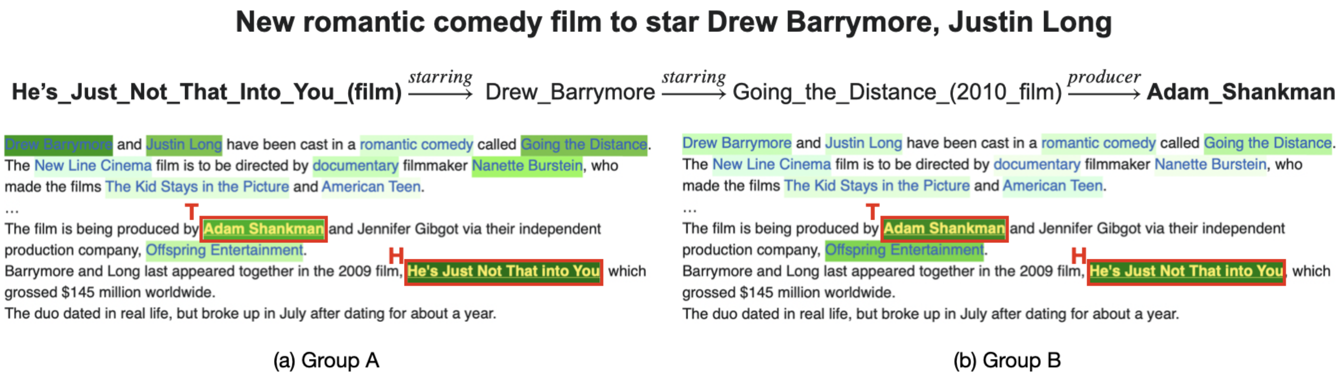
We identify features to represent both general and semantic user attention using the browsing behavior of a worker as he/she performs each assigned annotation task. The raw browsing behavior data of a worker at task, from which features will be extracted from, is a time series of browsing activities from the time the annotation task is displayed to the worker till the worker submits a "yes" or "no" answer.

Previous works determined that temporal, spatial, direction, speed, acceleration, clicks, descriptive statistics, and distribution features extracted from browsing data are useful

[6] The worker submits a "Yes" answer and the given knowledge path is indeed the ground truth contextual path, or the worker submits a "No" answer and the given knowledge path is not the ground truth contextual path.

[7] https://en.wikinews.org/wiki/New_romantic_comedy_film_to_star_Drew_Barrymore,_Justin_Long

[8] This might be due to the fact that we highlight both of these two mentions in the interface, as shown in Figure 1.



New romantic comedy film to star Drew Barrymore, Justin Long

He's_Just_Not_That_Into_You_(film) $\xrightarrow{starring}$ Drew_Barrymore $\xrightarrow{starring}$ Going_the_Distance_(2010_film) $\xrightarrow{producer}$ Adam_Shankman

Drew Barrymore and Justin Long have been cast in a romantic comedy called Going the Distance. The New Line Cinema film is to be directed by documentary filmmaker Nanette Burstein, who made the films The Kid Stays in the Picture and American Teen.
...
The film is being produced by Adam Shankman and Jennifer Gibgot via their independent production company, Offspring Entertainment.
Barrymore and Long last appeared together in the 2009 film, He's Just Not That into You, which grossed $145 million worldwide.
The duo dated in real life, but broke up in July after dating for about a year.

(a) Group A

Drew Barrymore and Justin Long have been cast in a romantic comedy called Going the Distance. The New Line Cinema film is to be directed by documentary filmmaker Nanette Burstein, who made the films The Kid Stays in the Picture and American Teen.
...
The film is being produced by Adam Shankman and Jennifer Gibgot via their independent production company, Offspring Entertainment.
Barrymore and Long last appeared together in the 2009 film, He's Just Not That into You, which grossed $145 million worldwide.
The duo dated in real life, but broke up in July after dating for about a year.

(b) Group B

FIGURE 2. Visualization of workers' attention on a Wikinews article. The head and tail entities are tagged with **H** and **T**.

TABLE 2. List of Features Used In Our Prediction Model

General Attention Features **(a) Base Features**	**General Attention Features** **(c) Focus Features**
Normalized viewpoint positions	Average time hovered on task entity mentions
Cursor normalized speed	Average time hovered on non-task entity mentions
Cursor normalized acceleration	Average time hovered on non-entity words
Cursor position status wrt. Wiki iframe	Average time hovered on task section
Distance traversed overall	Average time hovered on article section
x_{min}, x_{max},y_{min}, y_{max}, $\sigma_x(\sigma_y)$, $\mu_x(\mu_y)$	Average time hovered on Wiki iframe
Shannon entropy	Average time hovered on other elements
(Weighted) Permutation entropy	# highlighted task entity mentions
Approximate entropy	# highlighted non-task entity mentions
Fast Fourier Transformation	# highlighted non-entity words
General Attention Feature	Departure from shortest path (from last hovered element to submit)
(b) Segment Interaction Features	Time spent on reading the article (sec)
# Moves (towards, away) Wiki iframe	Time spent scrolling the Wiki iframe
# Moves (toward, away) Wiki iframe within dist. d	Time spent scrolling the article
# Clicks (inside, outside) Wiki iframe	Time spent for this annotation
Time to first click on Wiki iframe	**Semantic Attention Features**
Time to first hover on Wiki iframe	**(d)**
# Hovers over Wiki iframe	$z^w_{et} \times z^w_{\text{hv}}$
# Hovers over task section	$z^w_{dt} \times z^w_{\text{hv}}$
# Hovers over the other elements	$z^w_{d} \times z^w_{\text{hv}}$
# Hovers over Wiki iframe vs. other elements	$z^w_{et} \times z^w_{\text{at}}$
# Cursor positions within distance d from Wiki iframe	$z^w_{dt} \times z^w_{\text{at}}$
Distance traversed (inside, outside) Wiki iframe	$z^w_{d} \times z^w_{\text{at}}$
Distance traversed (inside, outside) task section	Note: z^w_{et}, z^w_{dt} and z^w_{d} are representations of
$\sum$ intra-distances of cursor positions wrt. Wiki iframe	task entities, news title and news article respectively.

in modeling user attention in search applications [36], [52], [53]. We consider these features to be relevant to the *general attention* of crowd-sourcing worker [30]. We categorize the general attention features into three sub-categories, namely: (a) base features, (b) segment interaction features, and (c) focus features, and adapt some of them to our annotation task. To capture the worker's attention of task semantics under *semantic attention factor*, we propose (d) semantic attention features. We show all these four categories of features in Table 2.

1) Base Features

Base features describe how a worker moves his/her mouse cursor. They include aggregated features extracted from cursor movement such as cursor speed, cursor acceleration, and descriptive statistics of the cursor positions. Specifically, the base features include four features: Shannon entropy, Permutation entropy, Approximate entropy, and Fast Fourier Transformation as shown in Table 2(a).

Shannon Entropy [54]. This aggregation feature characterizes the complexity of workers' mouse cursor trajectory. Shannon entropy provides a way to estimate the average minimum number of bits needed to encode a string of symbols in binary form based on the alphabet size and the frequency of the symbols. It has been shown to be effective in distinguishing engaged and non-engaged users [30].

Permutation Entropy [55]. We use both unweighted and weighted permutation entropies as features. Permutation entropy is another time series complexity measurement that considers the ordering between values and extract probability distribution of the ordinal patterns. The weighted permutation entropy (WPE) [56] is an extension of the permutation entropy. Unlike Permutation entropy, Weighted Permutation entropy further considers the amplitude information.

Approximate Entropy [57]. Approximate entropy summarizes the cursor trajectory's amount of regularity as well as its unpredictability of fluctuations. A high value of this entropy suggests more randomness in the cursor movement.

Fast Fourier Transform. Fast Fourier transform (FFT) is an efficient way of computing discrete Fourier transform. It is a spectral analysis that determines the frequency components in a time series. The frequency representation suggests how much of the variability of the data is caused by low or high frequencies. We use the ranking of frequency amplitude as features, e.g., first most powerful frequency, second most powerful frequency and so on.

2) Segment Interaction Features

These features capture the interactions the worker performs on different segments of an user interface, and the amount of attention given to these segments. In our crowdsourcing study, there are three segments, namely article section, task section, and Wiki iframe. Inspired by the idea that interaction with a specific segment of the webpage implies the level of worker's attention on that segment, we propose segment interaction features involving the external knowledge section (Wiki iframe) and the task section as shown in Table 2(b).

Table 2(b) includes some interaction features such as number of cursor moves towards and away from the Wiki iframe, number of clicks inside and outside Wiki iframe, number of hovers over Wiki iframe and number of hovers over task section. In addition, we introduce some interaction features involving the cursor positions from a specific segment within a distance parameter d. By limiting the observation within d, we are able to focus on interactions that are more likely to be related to the target element, the Wiki iframe in this case. In this study, we set $d = \frac{1}{5}W$ where W denotes the width of the screen size.

3) Focus Features

Here, we delve into the workers' attention paid on different parts of the article, and segments of the webpage. Consistent with previous works, we hypothesize user attention on certain content words or tokens during reading will affect annotation accuracy. We propose three types of content tokens: **(1) task entity mentions**: these are mentions of the two query entities of the task which are highlighted (in orange) in the annotation interface; **(2) non-task entity mentions**: these are mentions of other entities which are not highlighted but can be selected by the worker; **(3) non-entity words**: these are other words shown in the annotation interface. We extract focus features from the browsing behavior data on the above content token types including *time spent on hovering* and *number of mouse highlights on these tokens* as shown in Table 2(c).

The focus on task can also affect the time amount the worker spends on reading the article, scrolling through different sections, and annotating the task. We therefore measure time spent on different content tokens and on the annotation task as additional focus features. Finally, the deviation of cursor trajectory path from its last hovered position to the answer submission button position from the shortest cursor trajectory path between the two positions is also included as one of the focus features (i.e., *departure from shortest path* in Table 2). This feature was used in an earlier work to measure user's emotion status [39].

4) Semantic Attention Features

As discussed in Section V, more attention given to task-related entity mentions suggests higher annotation correctness. Again, attention paid to contextual path entity mentions, though has shown to be most effective in indicating annotation correctness, is hard to obtain in real-world crowdsourcing tasks. Thus, we instead examine how the worker's attention is semantically related to parts of the article that resemble contextual path entities.

To capture this semantic similarity, we represent the sequence of words and tokens the user has interacted with using embeddings. These features, also known as *worker attended content representations* shown in Table 2(d), include: **(a)** $z^w_{\mathbf{hv}}$: representation of word hovered **(b)** $z^w_{\mathbf{at}}$: representation of words that are paid more attention to (with hovered time > **t** threshold or highlighted). We use BERT [58] which provides contextualized embedding representations to encode the semantic of these word sequences.

As the name suggests, the contextual path carries information that is much dependent on the context. Thus, we choose to use the article title and the article itself to resemble the semantic of the contextual path entities. We encode news title, news article in the task using BERT to obtain their semantic representations. We also encode the task entity mentions in the same manner. These task content representations are denoted by z^w_{dt}, z^w_{d}, and z^w_{et} respectively. The dot product of different combinations of worker attended content representations and task content representations (i.e., $\{z^w_{\text{hv}}, z^w_{\text{at}}\} \times \{z^w_{et}, z^w_{dt}, z^w_{d}\}$) leads to six different features measuring the semantic similarity between the worker attended semantics and the task semantics.

B. EXPERIMENT SETUP

In this section, we evaluate the use of behavioral data to predict annotation quality of the workers. Using the 274 correct annotations as positive samples and the 216 incorrect annotations as negative samples, we train a logistic regression classifier with L1 penalty. In our experiments, we also try other classifiers but the results are similar to that of logistic regression. We show the average result over 10 rounds of 10-

TABLE 3. Prediction Performance Of Annotation Quality (RQ1 and RQ2)

Model	ACC	PRE	REC	F1
Baseline Models				
Random Baseline	50	56.1	49.2	52.4
Arapakis and Leiva, 2016 [30]	72.3	72.5	72.3	72.4
Yamauchi and Xiao [39]	62.5	62.6	62.2	62.4
Arapakis and Leiva, 2020 [29]	68.6	70.3	68.5	69.4
Models using General Attention Factor				
Base Features (**B**)	71.3	69.8	72.1	70.9
Segment Interaction Features (**S**)	68.6	69.5	68.3	68.9
Focus Features (**F**)	73.2	72.9	73.3	73.1
(B+S+F)	74.5	73.4	74.7	74
Model using Semantic Attention Features				
SemAtt	62.1	63.3	62.2	62.7
Model using All Features (Full)	**75.3**	**74.1**	**75.3**	**74.7**

TABLE 4. Top-20 Important Feature from The Full Model

Rk	Feature	Coef.	Cat
1	AVG time hovered on task section	3.182	F
2	Time spent for this annotation	3.117	F
3	AVG time hovered on task entity mentions	2.853	F
4	*Cursor normalized speed*	-2.441	B
5	# Hovers over task section	1.398	F
6	$z^w_{e^t}$ Representation of task entities $\cdot\ z^w_{\text{at}}$	1.132	S
7	Time spent on reading the article	1.072	F
8	*Distance traversed inside task section*	-1.071	I
9	*Shannon entropy*	-0.898	B
10	Distance traversed overall	0.728	B
11	# Hovers over Wiki iframe	0.561	F
12	z^w_d Representation of article $\cdot\ z^w_{\text{at}}$	0.552	S
13	# Hovers over task section	0.411	I
14	Σ_y	-0.259	B
15	*# Highlight non-entity words*	-0.236	F
16	AVG time hovered on Wiki iframe	0.194	F
17	*# Distance traversed outside task section*	-0.187	I
18	x_{min}	0.151	B
19	# Clicks inside Wiki iframe	0.132	I
20	*# Hovers over the other elements*	-0.094	I

Negatively correlated features are *underlined and italicized*.
(Feature Categories)
F: Focus, **B**: Base, **I**: Segment Interaction, **S**: Semantic

fold stratified cross validation such that the division of data samples into folds is different in each round. When using one fold of 49 samples as testing data, we train a prediction model on the remaining 441 samples. The performance metrics are then obtained from the prediction results for the testing data. The independent variables are all the extracted features (as shown in Table 2), and the prediction target is whether the annotation is correct or not.

To evaluate the different prediction models, we report the averaged accuracy (**ACC**), precision (**PRE**), recall (**REC**), and F1-score (**F1**) of the 10 folds in Table 3. We evaluate prediction models using different combinations of features:

- Prediction models using general attention feature sets only, i.e., **Model B** using base features, **Model S** using segment interaction features, **Model F** using focus features, and **Model B+S+F** using base, segment interaction, and focus features.
- Prediction model using semantic attention features **(SemAtt)**.
- Prediction model using all the above features **(Full)**

Moreover, we include a random baseline for comparison. The random model assigns each annotation task a label with a probability of 50% positive and 50% negative.

Last but not least, we also compare the performance using features from three previous user attention modeling works: Arapakis and Leiva, 2016 [30], Yamauchi and Xiao [39], and Arapakis and Leiva, 2020 [29]. We follow the same experiment setups as described in the papers. For the work by Arapakis and Levi [29], we use the time series encoder with GRU architecture to learn the trajectory representation for simplicity.

As shown in Table 3, all our proposed models outperform the random guess baseline. The best-performing model is the full model where all features are used. It achieves over 30% improvement in accuracy compared to the random baseline. Among models using different feature sets, the one using focus features achieves the highest accuracy, followed by base features. Although the model using semantic attention features does not perform well, it still outperforms the random baseline by roughly 15%.

We show the top-20 features of the trained full model in Table 4. Consistent with the performance of models using different feature sets, most of the top features are focus and base features. Generally, when a worker focuses more on task-related components (e.g., entity mentions, segment in interface), it will result in higher annotation accuracy. More detailed analysis of feature importance in Sections VI-C and VI-D.

C. GENERAL USER ATTENTION AND ANNOTATION QUALITY

Here, we address **Research Question 1**: "Are we able to establish the association between the workers' browsing behavior and their annotation task quality based on general user attention?" If the answer is affirmative, a corollary question is: "Which general attention features account most for the annotation task quality?" In this work, we adapt several browsing behavior features introduced in several previous works as different types of general user attention features. As our task quality prediction models are trained using these features, we can answer research question 1 based on these models' performance. Our experiments show that all prediction models using all general user attention features yield significantly higher accuracy than random baseline. We thus conclude that general attention features are clearly associated with worker's annotation quality.

We next examine baseline methods using different user attention features to predict annotation accuracy [29], [30], [39]. As shown in Table 3, the best performing baseline is the attention features proposed in [30] which achieve a 25% accuracy improvement compared to random guess baseline. The features used in this model have been covered under our base and segment interaction feature sets. The runner-up is the RNNs encoded trajectory proposed in [29] which only utilizes cursor trajectory feature to capture worker attention and it achieves around 20% accuracy improvement over random. This result suggests the importance of segment interaction features in the prediction of worker performance. The model from [39] which considers deviation of the worker's cursor trajectory from the shortest path to the answer submission button, on the other hand, shows only a relatively small accuracy improvement (14%) over random.

Beyond the attention feature sets found in the previous works, we examine the importance of three categories of general attention features introduced in this work, namely: base features, focus features, and segment interaction features. According to the result in Table 3, even the worst-performing model, i.e., prediction model using segment interaction features only, observes more than 20% improvement in accuracy over random baseline. Other single feature set-only prediction models also yield 23% to 25% accuracy improvement. By combining the three feature sets (i.e., prediction model using general attention factor features), we achieve 74.5% accuracy, which is 26% better than random baseline.

Table 4 also shows several general attention features assigned with large coefficients in the trained Full Model. These features are thus helpful in achieving accurate annotation task quality prediction. Most of the top ranked features are general user-attention features. Four out of five highest ranked feature are from the focus feature set, namely *Average time hovered on task section*, *Time spent for this annotation*, *Average time hovered on task entity mentions*, and *# Hovers over task section*. All these four features are positively correlated with good annotation quality, suggesting that higher accuracy may be a result of worker spending more time focusing on the annotation, task related segments in the annotation user interface, and parts of the article that are related to the task itself. *Cursor normalize speed* from base category, on the other hand, is a negatively correlated feature. This could be explained as a slower cursor movement might indicate higher annotation accuracy.

Based on the above findings, we conclude that general attention features contribute to the accuracy of annotation quality prediction.

D. SEMANTIC ATTENTION AND ANNOTATION QUALITY

Next, we address our **Research Question 2**: "Do semantic attention features contribute to annotation accuracy?" Different from Question 1 where substantial works has been proposed to capture general user attention, to the best of our knowledge, we have not found any work that explicitly models workers' semantic attention at the task level. We therefore propose our own semantic attention features as shown in Table 2.

We embed the sequence of words that are paid more attention by the worker, and compute the similarity between these words and the task information (i.e., entities, article title, and article content). As the embedding method we utilize in this work is based on contextual word embedding (i.e., BERT), the obtained vector representation also embeds some background information. Hence, the semantic similarity is not solely based on the literal definition of the words. Instead, it also implies how similar the two contexts are.

According to Tables 3 and 4, the model using only semantic attention features can only achieve 62.1% accuracy, which is a 14% improvement over random baseline. The semantic attention feature that contributes to the prediction most is $z_{et}^{w} \cdot z_{at}^{w}$, which represents the semantic similarity between task entities and words the worker has paid more attention to. This feature is positively correlated with annotation quality, suggesting that when a worker paid more attention to parts in the article that are semantically similar to the task entities, the annotation is more likely to be correct. Another important semantic attention feature with positive coefficient is, $z_{d}^{w} \cdot z_{at}^{w}$, which represents the semantic similarity between the whole article and words attended by the worker.

Although the prediction accuracy gain is less significant compared with other feature sets, semantic attention features still contribute to the overall prediction accuracy of the full model. Therefore, we are able to answer our second research question affirmatively as the prediction results back the effectiveness of the semantic attention features. In other words, we can conclude that by measuring the attention put to different parts of the article with semantic attention feature, the task-specific annotation quality can be predicted more accurately.

VII. EARLY PREDICTION OF THE ANNOTATION QUALITY

In this section, we address our **Research Question 3**: how soon can we predict the accuracy of an annotation? Is it possible to predict the annotation quality well before the annotation is completed by the worker? To address this question, we design experiments that make use of different proportions of browsing data to predict the annotation labels. Specifically, we segment an annotation's trajectory in two different ways.

The first way divides the trajectory data by actual time elapsed. On average, the workers spend 12.3 seconds to complete a task ($\sigma = 4.84$). 83.6% of the tasks are completed within 12.3 seconds. Thus, given an annotation trajectory, we sample the trajectory in the first 3, 6, 7, 9, 10, 11, 12, 13, 14, and 15 seconds, and build prediction models using the sampled trajectories. If the trajectory ends before the bin started (e.g., sampling the first 15 second of a task that finished within 10 seconds), we will simply use the whole trajectory for feature extraction.

The second way divides the trajectory by relative length of cursor movement. We build models with features extracted from 1/6, 2/6,..., 6/6 of the whole cursor movement path. This relative length approach is nevertheless less useful in applications as we are not able to know how long a cursor trajectory will be before the worker completes his/her annotation. This strategy therefore serves as a comparison with the division by time strategy in this study.

We use the same 274 correct and 216 incorrect annotations in our earlier experiments (see Section VI-B), and extract all features of the four feature sets. The prediction result is shown in Table 5. While the two trajectory division strategies are different, most of the workers completed their 2/6 of the whole trajectory within the first 6 seconds, and 5/6 of the trajectory within the first 12 seconds.

In the case of division by elapsed time strategy, we yields 60.6% accuracy using the data from the first 6 seconds, suggesting that early trajectory might not be very useful in determining the annotation quality. The accuracy however surges when the model is trained using data from the first 10 seconds. An accuracy of 72% could be achieved using the data from the first 12 seconds, which is already converging to the accuracy using the full data.

With the division by length strategy, we are able to achieve a 70.2% accuracy with 4/6 of the full trajectory data. However, even when using the data from the first 5/6 of the trajectory data, the accuracy is quite far from that obtain from full trajectory. One possible explanation is that the workers focus on the task section during the last part of the trajectory. According to the feature analysis in Table 4, features related to users' behavior in task section are quite important. Thus, by ignoring the last 1/6 of the trajectory, we also omit information crucial to the prediction accuracy.

This above result suggests that it is possible to predict the annotation quality before the annotation is done. However, there is still much room for the prediction accuracy to improve for this online prediction task. Furthermore, the extraction of features takes 2 to 3 seconds. More experiments need to be conducted to examine how this delay affects the prediction when one actually predict as the worker annotate. Moreover, our experiment is conducted using offline data. When observing workers in real time, we are not able to how long the whole cursor trajectory would be. Hence, the divide by length approach is invalid. This experiment is just to highlight the possibility of early-monitoring using workers' browsing data. The determination of when to confidently predict the annotation accuracy should be investigate in future works. The future work should also discuss how the early detection of worker's task-specific quality could be helpful for requester to design interventions. For instance, when the model detects a worker struggling with a task, the interface could show a message suggesting the worker to look for help or to skip the task. Either way, the requester can obtain a dataset of better quality.

VIII. DISCUSSION AND FUTURE WORK

While our experiment results demonstrate the connection between crowdsourcing workers' browsing behavior and their annotation quality, we further discuss additional findings, challenges, limitations, and future work of our study.

A. INTERACTION BETWEEN FRESH AND CONTINUED TASKS

A worker may perform multiple annotation tasks involving the same article during our crowdsourcing study. When this occurs, these tasks may or may not share the same query entities. Recall the first assigned task of the worker involving the same article is known to be the **fresh task**, while the subsequent task(s) involving the same article are known to be the **continued task(s)**. For the same article, we observe significant differences in some features between fresh tasks and continued tasks. For instance, *time spent on reading the article* and *time spent scrolling the article* is much longer in fresh tasks than in continued tasks. A possible explanation is that although we make the worker read the article thoroughly every time he/she receives a task by deactivating the submit button before he/she scrolls to the bottom of the article, the worker may not actually read the article again in the continued task(s). Instead, the worker may just focus on helpful part(s) of the article to complete the continued task. In other words, whether the worker focuses on reading the article in their fresh task may affect their annotation quality in the continued tasks.

To study the above carry-on effect from fresh tasks to the continued tasks, we conduct a t-test on the value of features from first and second tasks, and identify those with significant differences between the two tasks (i.e., p-value$\leq$ 0.01). Instead of deriving features from the browsing behavior of the task itself, we add some features from the fresh annotation task performed by a worker as additional features in the continued task(s) by the same worker involving the same article. These additional features in the continued tasks are known as *first encounter features* and are prefixed by $^{\mathrm{f}}$. For instance, given an article, for all tasks related to it (both fresh and continued tasks), the value of $^{\mathrm{f}}$*time spent on reading the article* is the value of *time spent on reading the article* of the corresponding fresh task. Thus, the behavior data from fresh task is passed to the continued tasks.

With these first encounter features, we train the annotation quality prediction model using the same training data in Section VI-B. The model yields 76.1% accuracy, 75.5% precision, 76.4% recall, and 75.9% F1, which is roughly 1% improvement over the model without first encounter features. We also conduct a feature analysis on this model as shown in Table 6. The top ranked first encounter features are $^{\mathrm{f}}$*Time spent on reading the article*, $^{\mathrm{f}}$*AVG time hovered on task entity mentions*, $^{\mathrm{f}}$*Time spent scrolling the article*, and $^{\mathrm{f}}$*AVG time hovered on article section*. All these features has positively coefficients, suggesting that if the worker has spent more time interacting with the article, especially the parts in the article that are related to the fresh task, the annotation quality might be higher.

Interestingly, we do not find the corresponding first encounter features of the top ranked features in Table 4 to be

TABLE 5. Online Prediction Result (RQ3)

Divide by Time				
Sec	**ACC**	**PRE**	**REC**	**F1**
3	53.4	60.2	56.3	58.2
6†	60.6	61.1	60.2	60.6
7	61.2	61.3	60.5	60.9
9	61.5	62.5	61.9	62.2
10	65.3	65.4	65.9	65.6
11	69.2	67.6	69.1	68.3
12⋆	72	70.1	71.9	71
13	73.7	72.6	72.3	72.4
14	74.4	73.5	74.2	73.8
15	74.7	74	74.9	74.4
all	75.3	74.1	75.3	74.7

Divide by Length				
Prop	**ACC**	**PRE**	**REC**	**F1**
1/6	57.4	58.1	58.8	58.4
2/6†	60.4	61	60.1	60.5
3/6	65.8	63.9	68.1	65.9
4/6	70.2	71.2	72.3	71.7
5/6⋆	73.3	72.8	74.1	73.4
all	75.3	74.1	75.3	74.7

† : On average, an annotator completed 2/6 of the browsing activities after 6 seconds
⋆ : On average, an annotator completed 5/6 of the browsing activities after 12 seconds

TABLE 6. Top-20 Important Feature from The Full Model (with First Encounter Features)

Rk	**Feature**	**Coef.**	**Cat.**
1	AVG time hovered on task section	3.229	F
2	$^{\text{f}}$**Time spent on reading the article**	2.711	F
3	$^{\text{f}}$**AVG time hovered on task entity mentions**	2.426	F
4	Time spent for this annotation	2.413	F
5	AVG time hovered on task entity mentions	2.192	F
6	*Cursor normalized speed*	-2.033	B
7	# Hovers over task section	1.552	F
8	z^w_{et} Representation of task entities · z^w_{at}	1.317	S
9	Time spent on reading the article	1.291	F
10	*Distance traversed (inside) task section*	-1.136	I
11	$^{\text{f}}$**Time spent scrolling the article**	0.915	F
12	*Shannon entropy*	-0.913	B
13	Distance traversed overall	0.769	B
14	$^{\text{f}}$**AVG time hovered on article section**	0.521	F
15	# Hovers over Wiki iframe	0. 512	F
16	z^w_d Representation of article · z^w_{at}	0.493	S
17	# Hovers over task section	0.272	I
18	Σ_y	-0.195	B
19	*# Highlight non-entity words*	-0.174	F
20	$^{\text{f}}$**# Hovers over task section**	0.097	F

Negatively correlated features are *underlined and italicized*.
First encounter features are in **bold** and with prefix $^{\text{f}}$.
(Feature Categories)
F: Focus, **B**: Base, **I**: Segment Interaction, **S**: Semantic

ranked highly in Table 6. One possible explanation is that such features are often related to how the worker work on the **current** task. For instance, *AVG time hovered on task section* only reflects how much time the worker actually spent on answering the task (i.e., reading the question then click on the answer button). This is somehow irrelevant to whether the worker will annotate the continued tasks correctly. Thus, its corresponding first encounter feature is not as important as itself in the prediction accuracy.

B. CHALLENGES AND LIMITATIONS

Based on our study results, we show that it is plausible to predict a worker's annotation quality based on his/her browsing behavior data. We justify why such data is helpful by introducing user attention model. Nevertheless, our results have not been able to establish a causal relationship between user attention and annotation quality in the crowdsourcing setting. Further research and user studies should be conducted to establish this formal theoretical connection. It is also interesting to identify any other important factors to be considered and also their respective browsing behavior features for training even more accurate task-quality prediction models.

As we conducted our study on AMT workers, it was difficult to control the workers' annotation environment. The AMT workers could come from any parts of the world, using devices of difference screen sizes. To collect a set of browsiour behavior data less affected by device choices for analysis, our study required workers to use desktop or laptop only. They are also not allowed to resize the window to prevent resizing noises in the extracted features. In other words, our study rely heavily on good cursor movement data gathering, which may not work well on mobile devices with touch screens such as smart phones or tablets. To generalize this research to mobile devices, one may consider employing one of the eye-tracking systems that utilize the mobile built-in camera to capture user attention [59], [60]. One may need to refer to user attention works based on eye-tracking to construct features for modeling worker annotation quality on mobile devices.

Finally, in this study we do not propose features that work for all kinds of tasks, instead; we show the predictive value of user attention modeling features in annotation quality. Hence, many of our features (e.g., most of our segment interaction, focus, and semantic attention feature) are specific to information retrieval problems similar to contextual path retrieval. More research should be conducted on other interesting application problems with different user attention requirements.

C. FUTURE WORK

As part of future research, we plan to focus on modeling individual variations as workers perform the annotation tasks. Studies have shown that there are individual differences in cursor moving pattern when users perform some web search tasks [33]. Such individual deviations may be captured as user-level behavior features in addition to the task-specific behavior features used in our work. In addition, task difficulty and knowledge required to complete tasks should also be taken into consideration when predicting annotation task quality. Finally, with significant advances in behavior data representations using neural networks [29], our work can be extended to represent a cursor path using a vector representation instead of descriptive features so as to capture salient behavioral semantics that improve task-specific quality prediction.

There several directions to extend our work to other information retrieval problems. First, besides user-attention models, there may be other models that can be used to predict worker's task-specific quality. For example, reading comprehension is a type of question answering task where the workers are to read an article, then answer several questions about it. The reading behaviors indicative of content understanding, such as reading speed, may affect the annotation quality and should be considered in the study. Through analysing browsing behavior of several type annotation tasks, it is then possible to generalize a set of factors that predicts task quality with robust accuracy.

Early task quality prediction can be used to intervene worker annotation to achieve better crowdsourcing outcome. Other than determining if an annotation is good or bad, it can be used to assign the right tasks to the right workers. With appropriate annotation interface design, we could also determine how a worker can be engaged at the right moment to improve his/her annotation quality. Some previous works suggest that worker produce annotation of better quality when exposed to some form of supervision [61]. As our early prediction model predicts the workers' performance in real time, it is possible to use the prediction result to give the worker a sense of being monitored.

IX. CONCLUSION

In this work, we investigate into the connection between crowdsourcing workers' annotation quality and their behavior. We postulate the connection between user attention and annotation quality and model user attention as features that are subsequently used for annotation quality prediction. We conduct qualitative and quantitative analysis on how the workers focus on different parts of the article and how it affects the annotation correctness. We propose semantic user attention features based on these finding. In addition, we propose general user attention and semantic user attention that covers non-task semantics. Our experiments on behavior data collected from specially instrumented annotation user interface show better accuracy in the prediction results than state-of-the-art and baseline models using our proposed attention features.

ACKNOWLEDGMENT

This research is partially supported by the Lee Kong Chian Professorship and National Research Foundation, Singapore under its Strategic Capabilities Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

REFERENCES

[1] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In EMNLP, 2008.

[2] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In CIKM, 2012.

[3] Lilly C. Irani and M. Six Silberman. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In CHI, 2013.

[4] Masatomo Kobayashi, Shoma Arita, Toshinari Itoko, Shin Saito, and Hironobu Takagi. Motivating multi-generational crowd workers in social-purpose work. In CSCW, 2015.

[5] Alexey Tarasov, Sarah Jane Delany, and Brian Mac Namee. Dynamic estimation of worker reliability in crowdsourcing for regression tasks: Making it work. Expert Systems with Applications, 41(14):6190–6210, 2014.

[6] David R Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. Operations Research, 62(1):1–24, 2014.

[7] Alex Williams, Joslin Goh, Charlie Willis, Aaron Ellison, James Brusuelas, Charles Davis, and Edith Law. Deja vu: Characterizing worker reliability using task consistency. In HCOMP, 2017.

[8] Chien-Ju Ho and Jennifer Vaughan. Online task assignment in crowdsourcing markets. In AAAI, 2012.

[9] Leyla Kazemi, Cyrus Shahabi, and Lei Chen. Geotrucrowd: Trustworthy query answering with spatial crowdsourcing. In SIGSPATIAL, 2013.

[10] Ju Fan, Guoliang Li, Beng Chin Ooi, Kian-lee Tan, and Jianhua Feng. Icrowd: An adaptive crowdsourcing framework. In SIGMOD, 2015.

[11] Robin Wentao Ouyang, Lance Kaplan, Paul Martin, Alice Toniolo, Mani Srivastava, and Timothy J. Norman. Debiasing crowdsourced quantitative characteristics in local businesses and services. In IPSN, 2015.

[12] Jeff Huang, Ryen White, and Georg Buscher. User see, user point: gaze and cursor alignment in web search. In CHI, 2012.

[13] Vidhya Navalpakkam, LaDawn Jentzsch, Rory Sayres, Sujith Ravi, Amr Ahmed, and Alex Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In WWW, 2013.

[14] Ricky KP Mok, Rocky KC Chang, and Weichao Li. Detecting low-quality workers in qoe crowdtesting: A worker behavior-based approach. IEEE Transactions on Multimedia, 19(3):530–543, 2016.

[15] Jyun-Cheng Wang and Rong-Fuh Day. The effects of attention inertia on advertisements on the www. Computers in Human Behavior, 23(3):1390–1407, 2007.

[16] Yu-Chen Hsieh and Kuo-Hsiang Chen. How different information types affect viewer's attention on internet advertising. Computers in human Behavior, 27(2):935–945, 2011.

[17] Qiuzhen Wang, Sa Yang, Manlu Liu, Zike Cao, and Qingguo Ma. An eye-tracking study of website complexity from cognitive load perspective. Decision support systems, 62:1–10, 2014.

[18] Florian Mueller and Andrea Lockerd. Cheese: tracking mouse movement activity on websites, a tool for user modeling. In CHI Extended Abstracts on Human Factors in Computing Systems, 2001.

[19] Mark Claypool, Phong Le, Makoto Wased, and David Brown. Implicit interest indicators. In IUI, 2001.

[20] Bracha Shapira, Meirav Taieb-Maimon, and Anny Moskowitz. Study of the usefulness of known and new implicit indicators and their optimal combination for accurate inference of users interests. In SAC, 2006.

[21] Luis A Leiva and Jeff Huang. Building a better mousetrap: Compressing mouse cursor activity for web analytics. Information Processing & Management, 51(2):114–129, 2015.

[22] Dmitry Lagun and Eugene Agichtein. Inferring searcher attention by jointly modeling user interactions and content salience. In SIGIR, 2015.

[23] Daniel Martín-Albo, Luis A Leiva, Jeff Huang, and Réjean Plamondon. Strokes of insight: User intent detection and kinematic compression of mouse cursor trails. Information Processing & Management, 52(6):989–1003, 2016.

[24] Dmitry Lagun, Mikhail Ageev, Qi Guo, and Eugene Agichtein. Discovering common motifs in cursor movement data for improving web search. In WSDM, 2014.

[25] Ye Chen, Yiqun Liu, Min Zhang, and Shaoping Ma. User satisfaction prediction with mouse movement information in heterogeneous search environment. TKDE, 29(11):2470–2483, 2017.

[26] Qi Guo and Eugene Agichtein. Exploring mouse movements for inferring query intent. In SIGIR, 2008.

[27] Qi Guo and Eugene Agichtein. Ready to buy or just browsing? detecting web searcher goals from interaction data. In SIGIR, 2010.

[28] Ioannis Arapakis, Mounia Lalmas, and George Valkanas. Understanding within-content engagement through pattern analysis of mouse gestures. In CIKM, 2014.

[29] Ioannis Arapakis and Luis A Leiva. Learning efficient representations of mouse movements to predict user attention. In SIGIR, 2020.

[30] Ioannis Arapakis and Luis A Leiva. Predicting user engagement with direct displays using mouse cursor information. In SIGIR, 2016.

[31] Mikhail Ageev, Dmitry Lagun, and Eugene Agichtein. Improving search result summaries by using searcher behavior data. In SIGIR, 2013.

[32] Jeff Huang, Ryen W White, Georg Buscher, and Kuansan Wang. Improving searcher models using mouse cursor activity. In SIGIR, 2012.

[33] Kerry Rodden, Xin Fu, Anne Aula, and Ian Spiro. Eye-mouse coordination patterns on web search results pages. In CHI Extended Abstracts on Human Factors in Computing Systems, 2008.

[34] Daniel J Liebling and Susan T Dumais. Gaze and mouse coordination in everyday work. In UbiComp, 2014.

[35] David Hauger, Alexandros Paramythis, and Stephan Weibelzahl. Using browser interaction data to determine page reading behavior. In UMAP, 2011.

[36] Vidhya Navalpakkam and Elizabeth Churchill. Mouse tracking: measuring and predicting users' experience of web-based content. In CHI, 2012.

[37] Takashi Yamauchi and Casady Bowman. Mining cursor motions to find the gender, experience, and feelings of computer users. In ICDM, 2014.

[38] Yiqun Liu, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, and Xuan Zhu. Different users, different opinions: Predicting search satisfaction with mouse movement information. In SIGIR, 2015.

[39] Takashi Yamauchi and Kunchen Xiao. Reading emotion from mouse cursor motions: Affective computing approach. Cognitive science, 42(3):771–819, 2018.

[40] Eli T Brown, Alvitta Ottley, Helen Zhao, Quan Lin, Richard Souvenir, Alex Endert, and Remco Chang. Finding waldo: Learning about users from their interactions. TVCG, 20(12):1663–1672, 2014.

[41] Eugene Yujun Fu, Tiffany C.K. Kwok, Erin You Wu, Hong Va Leong, Grace Ngai, and Stephen C.F. Chan. Your mouse reveals your next activity: Towards predicting user intention from mouse interaction. In COMPSAC, 2017.

[42] Tiffany C.K. Kwok, Eugene Yujun Fu, Erin You Wu, Michael Xuelin Huang, Grace Ngai, and Hong-Va Leong. Every little movement has a meaning of its own: Using past mouse movements to predict the next interaction. In IUI, 2018.

[43] Jeffrey M. Rzeszotarski and Aniket Kittur. Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In UIST, 2011.

[44] Shuguang Han, Peng Dai, Praveen Paritosh, and David Huynh. Crowdsourcing human annotation on web page structure: Infrastructure design and behavior-based quality control. ACM Trans. Intell. Syst. Technol., 7(4):1–25, 2016.

[45] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. Strategies for crowdsourcing social data analysis. In CHI, 2012.

[46] Jingzheng Tu, Guoxian Yu, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. Attention-aware answers of the crowd. In SDM, 2020.

[47] Ujwal Gadiraju and Gianluca Demartini. Understanding worker moods and reactions to rejection in crowdsourcing. In HT, 2019.

[48] Andrew Mao, Ece Kamar, and Eric Horvitz. Why stop now? predicting worker engagement in online crowdsourcing. In HCOMP, volume 1, 2013.

[49] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In ICWSM, 2011.

[50] Steve Branson, Catherine Wah, Florian Schroff, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie. Visual recognition with humans in the loop. In ECCV, 2010.

[51] Gabriella Kazai. In search of quality in crowdsourcing for search engine evaluation. In ECIR, 2011.

[52] Qi Guo, Dmitry Lagun, Denis Savenkov, and Qiaoling Liu. Improving relevance prediction by addressing biases and sparsity in web search click data. In WSDM, 2012.

[53] Qi Guo and Eugene Agichtein. Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior. In WWW, 2012.

[54] Claude Elwood Shannon. A mathematical theory of communication. ACM SIGMOBILE, 5(1):3–55, 2001.

[55] Christoph Bandt and Bernd Pompe. Permutation entropy: a natural complexity measure for time series. Physical review letters, 88(17):174102, 2002.

[56] Bilal Fadlallah, Badong Chen, Andreas Keil, and José Príncipe. Weighted-permutation entropy: A complexity measure for time series incorporating amplitude information. Physical Review E, 87(2):022911, 2013.

[57] Steven M Pincus. Approximate entropy as a measure of system complexity. PNAS, 88(6):2297–2301, 1991.

[58] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL, 2018.

[59] Nachiappan Valliappan, Na Dai, Ethan Steinberg, Junfeng He, Kantwon Rogers, Venky Ramachandran, Pingmei Xu, Mina Shojaeizadeh, Li Guo, Kai Kohlhoff, et al. Accelerating eye movement research via accurate and affordable smartphone eye tracking. Nature communications, 11(1):1–12, 2020.

[60] Lucas Paletta, Helmut Neuschmied, Michael Schwarz, Gerald Lodron, Martin Pszeida, Stefan Ladstätter, and Patrick Luley. Smartphone eye tracking toolbox: accurate gaze recovery on mobile displays. In ETRA, 2014.

[61] Babak Naderi, Ina Wechsung, and Sebastian Möller. Effect of being observed on the reliability of responses in crowdsourcing micro-task platforms. In QoMEX, 2015.

PEI-CHI LO received her B.S. and M.B.A. degree in Information Management from National Sun Yat-sen University, Kaohsiung, Taiwan. She is currently pursuing the Ph.D. degree in Computer Science at Singapore Management University, Singapore.In 2017, she was a research engineer in Living Analytic Research Centre, Singapore. Her research interest includes knowledge graph-based information retrieval and representation learning, behavior-based user modeling in crowdsourcing, and computational linguistics.

EE-PENG LIM is the Lee Kong Chian Professor with the School of Computing and Information Systems at the Singapore Management University (SMU). Dr Lim received his PhD degree from University of Minnesota. His research expertise covers social media mining, social/urban data analytics, and information retrieval. He has published more than 400 international journal and conference papers from his research works. He is the recipient of the Distinguished Contribution Award at the 2019 Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD) and Test of Time Award at the 2020 ACM International Conference of Web Search and Data Mining (WSDM). He currently serves on Singapore's Social Science Research Council which focuses on developing talent and strengthening social science and humanities research that benefits social and economic development in Singapore and the Asian region.