Human Behavior Inspired Machine Reading Comprehension

Yukun Zheng†, Jiaxin Mao†, Yiqun Liu†*, Zixin Ye‡, Min Zhang†, and Shaoping Ma†
†Department of Computer Science and Technology, Institute for Artificial Intelligence,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
‡School of Computer Science and Engineering, Beihang University, Beijing 100083, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

Machine Reading Comprehension (MRC) is one of the most challenging tasks in both NLP and IR researches. Recently, a number of deep neural models have been successfully adopted to some simplified MRC task settings, whose performances were close to or even better than human beings. However, these models still have large performance gaps with human beings in more practical settings, such as MS MARCO and DuReader datasets. Although there are many works studying human reading behavior, the behavior patterns in complex reading comprehension scenarios remain under-investigated. We believe that a better understanding of how human reads and allocates their attention during reading comprehension processes can help improve the performance of MRC tasks. In this paper, we conduct a lab study to investigate human's reading behavior patterns during reading comprehension tasks, where 32 users are recruited to take 60 distinct tasks. By analyzing the collected eye-tracking data and answers from participants, we propose a two-stage reading behavior model, in which the first stage is to search for possible answer candidates and the second stage is to generate the final answer through a comparison and verification process. We also find that human's attention distribution is affected by both question-dependent factors (e.g., answer and soft matching signal with questions) and question-independent factors (e.g., position, IDF and Part-of-Speech tags of words). We extract features derived from the two-stage reading behavior model to predict human's attention signals during reading comprehension, which significantly improves performance in the MRC task. Findings in our work may bring insight into the understanding of human reading and information seeking processes, and help the machine to better meet users' information needs.

KEYWORDS

Reading comprehension, reading behavior model, user behavior analysis, eye tracking

*Corresponding author.

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07.
https://doi.org/10.1145/3331184.3331231

1 INTRODUCTION

Although a number of machine reading comprehension (MRC) models [33, 36, 37, 41] have shown the ability to approach or even transcend human beings in a few simplified MRC tasks like SQuAD [26], they are still far behind human beings in a number of challenging MRC tasks, such as MS MARCO [23] and DuReader [14]. Therefore, we consider that it is necessary to study how human accomplishes such reading comprehension tasks. Better understanding human's reading behavior patterns may inspire MRC models to achieve better performance.

Reading is one of the most fundamental channels of gaining knowledge and information, during which human simultaneously processes visual signals and perceives information. Plenty of research [6, 7, 15, 17, 28, 29] has studied the human reading behavior in the past few decades and proposed a number of human reading behavior models. According to the reading context settings, we classify the proposed human reading models into two categories: general reading models and specific reading models under a certain context. The first category includes E-Z model [28, 29], SWIFT [9, 10] and the Bayesian reading model [3], which formalized the human reading patterns in non-contextual reading settings. The second category includes Two-Stage Examination Model [20], Reading Model in Relevance Judgment [18], which were proposed to respectively model the examination behavior on search engine result pages and the reading behavior patterns during relevance judgment process. In a reading comprehension task, human reads with clear intent, i.e., to find the most appropriate answer to the question, which may influence the reading behavior. Therefore, the general reading models or reading models under other contexts may be inapplicable in this kind of reading comprehension scenario. However, the behavior patterns of human in such reading comprehension scenario remain under-investigated, which motivates our first research question:

- **RQ1**: How do humans read and seek answers during reading comprehension tasks?

A number of works [15, 28, 28, 29] proved that eye tracking is an effective measure to understand the state of human cognitive behavior. Two kinds of eye movements are usually used in analyzing human reading patterns: fixation and saccade [29]. In this paper, we mainly focus on the eye fixation behavior to study the reading patterns, especially the attention allocation mechanism, during reading comprehension. Reichle et al. [29] proposed that during reading, the fixation location is jointly determined by linguistic, visual, and oculomotor factors, while the fixation duration is decided by ongoing linguistic processing. As we know, the factors that may affect human attention include but are not limited to position [18],

word frequency [27], predictability in the context [1] and Part of Speech [2]. In reading comprehension tasks, besides these general factors, the attention may also be affected by question-dependent factors, such as answer location and matching signals between the question and the document. To verify the effects of these factors, we aim to study the second research questions:

- **RQ2**: What factors affect the attention allocation mechanism of human beings during reading comprehension tasks?

Human behavior, especially eye tracking, has shown the effectiveness in many Natural Language Processing (NLP) and Information Retrieval (IR) tasks, such as Part-of-Speech Tagging [2], machine translation error analysis [32], named entity annotation [34], click prediction [21] and relevance judgment [4]. However, there is still a lack of research on the effectiveness of human behavior in MRC tasks. Therefore, by investigating the attention allocation mechanism of human beings, we propose our third research question:

- **RQ3**: Can learning from features of human behavior (e.g., eye tracking) help improve the performance of MRC tasks?

In this study, we conduct a lab-based user study to collect user behavior data and various annotations in reading comprehension tasks. User behavior data is collected in the processes of reading both the question and document, including eye tracking and mouse movements. After reading, users are asked to answer the question and then highlight the evidence snippets which fully support the answer they write. Based on the collected data, we investigate the reading behavior patterns of humans by dividing the reading process into equal time periods and compare the reading behavior between with-answer documents and without-answer ones. When analyzing the effects of question dependent and independent factors on human attention, we use analysis of variance (ANOVA) to determine the effect size of these factors. To verify the effectiveness of human attention in the MRC task, we try two approaches: 1) directly feed real human attention into answer retrieval models; 2) train an attention prediction model and use the predicted attention as features in answer retrieval. The main contributions of this paper are listed as follows:

- We conduct a lab-based user study to collect various kinds of user behavior data and annotations in reading comprehension tasks, which will be released to the research community.
- We first propose a two-stage reading behavior model and reveal the effects of question dependent and independent factors on human attention in reading comprehension tasks.
- By incorporating real and predicted human attention into answer retrieval models, we show its effectiveness in improving the performance of MRC tasks.

The remainder of this paper is organized as follows. We review related works on reading behavior models and MRC models in Section 2 and describe the details of our user study in Section 3. In Section 4, we analyze the user behavior data to addresses **RQ1**. To answer **RQ2**, we investigate the effects of question-dependent and question independent factors on human attention in Section 5. We design a two-step answer prediction model to investigate **RQ3** in Section 6 and finally conclude our work in Section 7.

2 RELATED WORK

Reading is a complicated physiological and psychological process involving vision processing, language understanding, information gaining, nerves controlling [7]. Understanding how human reads is a challenging problem and has been studied for decades in various fields, such as psychology, linguistics, computer science and neurology. Here we briefly introduce several influential reading models, which can be classified into two categories: general reading models and specific reading models under a certain context.

General reading models. Crowder and Wagner [7] developed a bilateral cooperative model of reading from the psychologist's perspective, which covers from letter recognition to reading whole texts. They assumed that there are two simultaneous track during human reading: one dealing with linguistics and another relating the meanings of words and phrases with real-world conditions. Reichle et al. [29] proposed an EZ Reader model, which provided a comprehensive account of eye movement control during reading, i.e., how word identification, visual processing, attention, and oculomotor control jointly determine when and where the eyes move. Just and Carpenter [15] presented a reading model on the allocation mechanism of eye fixations during reading and found that readers make longer pauses when their processing loads are greater, such as accessing infrequent words.

Specific reading models. Liu et al. [20] proposed a two-stage examination model on the examination behavior on search engine result pages (SERPs) and found that there usually exists a skimming step before the user really reads the search result, which can help estimate better relevance of search results. Li et al. [18] found that during the relevance judgment process, there are significant differences in the human attention distribution between relevant and irrelevant documents. Based on the findings of eye tracking data, they proposed a two-stage reading behavior model in the relevance judgment process, which consists of preliminary relevance judgment stage and reading with preliminary relevance stage.

Eye tracking, especially fixation and saccade, has been widely studied in decades. A fixation is the maintaining of the visual gaze on a single location, which typically lasts a brief period (200-250 ms), while a saccade is a jump carrying the eye from one fixation to another, which typically lasts 20-50 ms [29]. In this work, we focus on the fixation behavior and consider it as a kind of human attention in reading comprehension tasks. Existing works [1, 2, 18, 27] showed that human attention during reading is usually affected by many factors. Li et al. [18] revealed that human attention is strongly biased by vertical position during relevance judgment process. Rayner [27] found that readers look longer at low-frequency words than at high-frequency words. Altarriba et al. [1] found that words that are highly predictable from the preceding context will be fixated for less time than lowly predictable words. Eye gaze patterns during reading were found to be strongly affected by the Part-of-Speech of words [2]. A number of works [2, 4, 21, 32, 34] used eye tracking to improve the performances in NLP and IR tasks. Barrett et al. [2] took advantage of eye-tracking data to improve the performance of weakly supervised models in the Part-of-Speech Tagging task. Stymne et al. [32] proposed to use eye tracking to enhance baseline approaches in the error analysis of machine translation. Liu et al. [21] studied the users' examination behavior in different click sequences based on eye tracking and incorporated

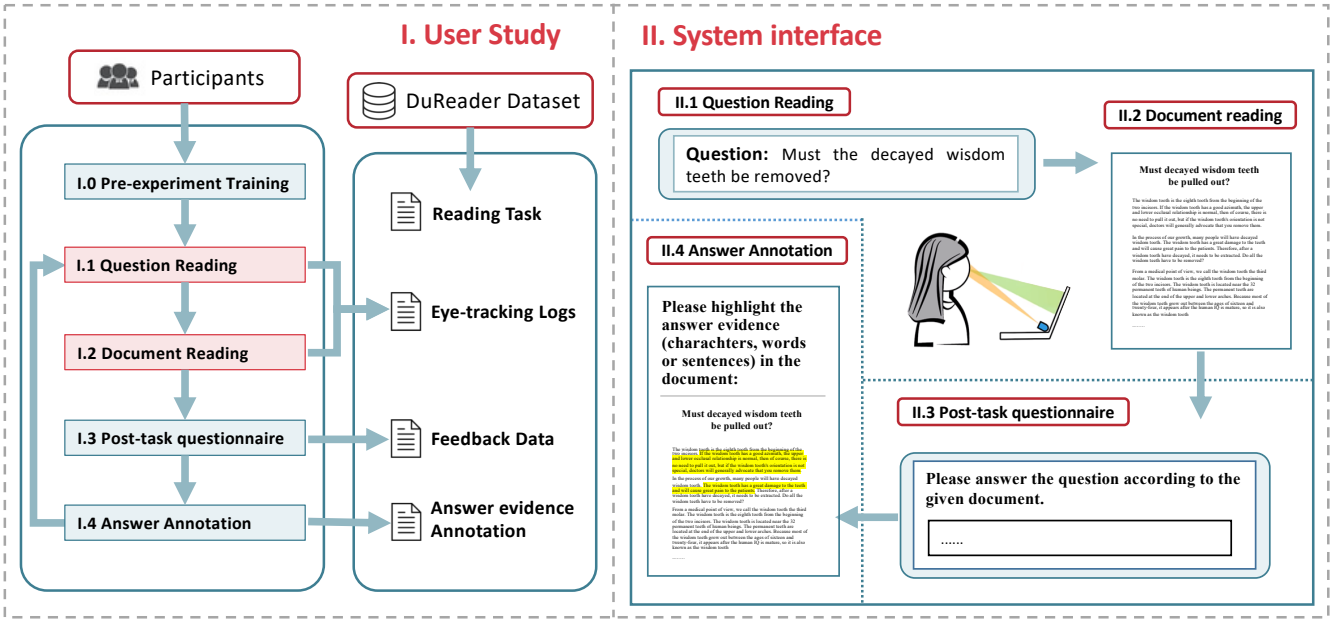


Figure 1: The procedure and system interface of the user study.

their findings into Time-Aware Click Model, which outperformed all baselines in the click prediction task.

Recently, several neural MRC models [8, 30, 31, 35, 37, 38] have been proposed with the inspiration from human reading behavior. Inspired by the reread behavior of humans, multi-turn reasoning mechanism has been introduced into MRC models to improve the performance [8, 30, 31, 35, 38]. Wang et al. [37] proposed an efficient mechanism of cross-passage answer verification to address the challenge that there are multiple candidate answers in passages. However, both multi-turn reasoning and answer verification mechanisms are introduced based on intuition or experiences rather than reliable observations of human behavior in the reading comprehension tasks. Therefore, we would like to investigate whether these mechanisms can be observed in our study.

3 USER STUDY

3.1 Tasks and Participants

Among several popular MRC datasets, we choose the DuReader[14] dataset, one of the most popular and challenging Chinese MRC dataset, because of its practical settings. We select 15 questions from DuReader, whose types include *Description*, *Entity* and *Yesno* (5 questions for each), and cover both *Fact* (8 questions) and *Opinion* (7 questions) categories. For the five documents of a question, we filter out those documents where the number of paragraphs is less than 3 or more than 20. Finally, 4 documents are kept per question. Most of the documents provide more than one precise answer to the question, but there exist a few documents with no answer. We divide the documents into 4 groups. Each group covers all 15 questions and consists of 15 distinct documents of these questions. During the user study, the annotation system randomly shows the 15 tasks of a certain group to the participant one by one.

We recruit 32 participants (also called users in the following paper) to take our reading comprehension tasks. There are 21 males and 11 females with their ages ranging from 18 to 26. All of them are undergraduate or graduate students and their majors vary from natural science and engineering to humanities and sociology. All participants possess college-level skills in Chinese reading comprehension and skillful computer operation capability. In addition, we screen all applicants according to their visual acuity to ensure that the collected eye movements are correct. It takes about one and a half to two hours to accomplish 15 tasks and each participant is paid about $15.

3.2 Procedure

The procedure and system interface of our user study are shown in Figure 1. In the beginning, participants are tutored to accomplish an example task as the pre-experiment training. Next, they are required to finish 15 reading comprehension tasks independently with a 15-minute break halfway. The procedure of a complete reading comprehension task is as follows:

Question reading. Participants are shown a question in the center of the screen and they need to read and memorize it. At the same time, the eye tracker collects the eye movements of participants. After reading, they need to rewrite the question to ensure that they really remembered it. They have two chances to return to the former page and reread the question. Then, participants need to finish a pre-task questionnaire about the difficulty, interest and understanding level of the question on a five-level scale. In the next document-reading step, the question will not be presented.

Document reading. One corresponding document of the question is presented to participants. They need to read the document and find the best answer to the question in the documents. Their

Table 1: The statistics of our user study.

#Questions	#Docs	#Users	#Sessions	#Valid sessions
15	60	32	480	406

eye movements are collected by the eye tracker during reading, while the mouse movements are also recorded by our system.

Post-task questionnaire. After reading the document, participants are asked to write down the answer in the post-task questionnaire. The answers are required to be as precise and complete as possible and must come straight from the document or be supported by evidence snippets in the document instead of being generated based on the participant's own knowledge. Since a few documents contain no answer, participants should just answer "None" if they consider there is no answer in the document. There are also two chances for them to return to the former page and reread the document in case that they forget some details of the answer. Besides collecting the answer, we also ask the difficulty, interest and understanding level of the question, the readability, relevance and usefulness of the document, the answer findability and quality (correctness, completeness and conciseness). Each annotation is on a five-level scale.

Answer annotation. In this step, the question and the document are presented to participants again and they are asked to highlight the snippets of answer evidence (any words, phrases or sentences) in the document, which must be able to fully support their answers. If the document cannot provide an answer, they don't need to highlight anything. Then they should label the usefulness of each paragraph in the document with respect to the question on a four-level scale (4: extremely useful; 3: fairly useful; 2: somewhat useful; 1: useless).

3.3 Collecting User Behavior

We collect two kinds of user behavior data in the user study: eye-tracking and mouse movements. For the eye-tracking data, we use a Tobii X2-30 eye tracker to record the eye movements of participants during reading questions and documents, whose deviation is within the word level. Before taking tasks, there is a calibration process for each participant to ensure that the data of eye movements can be recorded accurately. During reading questions and documents, there is only a question or document at one time without any other components. Participants need to press the space bar to enter the next page when they finish reading, which is designed to reduce the interference of human-computer interaction to the collected eye movement data. For the data of mouse movements, we collect three types of mouse events: *move*, *scroll* and *select*. For the mouse moving and scrolling events, the mouse coordinates at the event moments are saved in our user behavior log. For the selecting events, we record the corresponding selected words and their positions in the documents.

3.4 Collected Data

Table 1 shows the statistics of the user study. With 32 valid participants, we collect 480 sessions of 60 reading comprehension tasks in total [1]. Each document has been read by eight individual participants. Besides the eye tracking and mouse movements, we also collect the rewritten questions, answers written by participants (user answers) and highlighted snippets of answer evidence (annotated answers). Based on these data, we can examine whether the data of a session is reliable. We invite an external expert annotator to review all sessions and filter out sessions where the rewritten question and the user answer are wrong or the annotated answer is inconsistent with the corresponding user answer. Finally, 406 valid sessions of 59 tasks remain where 66 sessions of 10 tasks are no-answer. On average, a question and a document have been read 1.02 times and 1.15 times respectively in a session. We incorporate the user behavior data of multiple reads in a session into a whole in the following experiment.

To measure the agreement of different kinds of answers, we calculate Rouge-L [19] on the valid sessions using the evaluation script [2] of DuReader. For the annotated answers, every time we choose one participant's answer as the candidate answer and others' answers in the same task as the reference answers. The average Rouge-L of annotated answers on the whole tasks is 89.72. The same Rouge-L of user answers is 77.94. The average Rouge-L between user answers and the annotated answers in a session is 62.86, indicating that the answers have been rewritten based on the answer evidence in the documents.

The Fleiss' κ [11] is used to measure the inner-person agreement of usefulness annotations, which ranges from 0 to 1 (0-0.2: slight agreement, 0.2-0.4: fair agreement, 0.4-0.6: moderate agreement, 0.6-0.8: substantial agreement, 0.8-1.0: almost perfect agreement [16]). The Fleiss' κ of 4-level passage usefulness from eight participants is 0.450, reaching a moderate agreement level. At the sentence level, if the words of a sentence have been highlighted as the answer evidence, the usefulness of this sentence is labeled as 1. Otherwise, the sentence's usefulness is 0. According to the statistics, there are 2.1 useful sentences in one session on average and 53.3% sessions have more than one useful sentence. The Fleiss' κ of binary sentence usefulness from eight participants is 0.537, also reaching a moderate agreement level.

All these statistics show that the data of answers and annotations we collected in the user study is reliable to be used in the following analysis and experiment.

4 READING BEHAVIOR MODEL

To address **RQ1**, we make a deep analysis of the user study data in this section and propose a two-stage reading behavior model in reading comprehension.

4.1 Answer Seeking

We first investigate the reading order of users. Figure 2 shows the first arrival time at five-level vertical positions and the distribution of reading time on documents. We can see that the average first arrival time increases with the vertical positions. The average time of reading documents is 98.4 seconds and the average first arrival time at the bottom (80%-100%) of documents is 58.5 seconds, which approximately occurs at the 60% moment of the reading process.

[1] The data of our user study, including tasks, user behavior data and annotations, will be released after the double-blind review.
[2] https://github.com/baidu/DuReader/blob/master/utils/dureader_eval.py

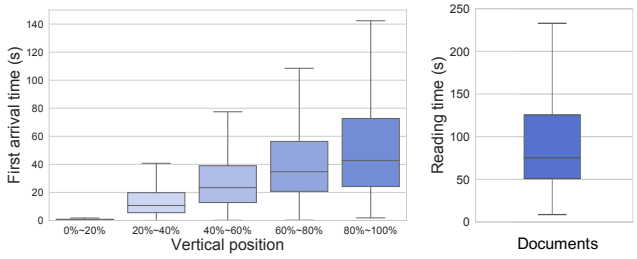


Figure 2: The first arrival time at five-level vertical positions and the distribution of reading time on documents.

Figure 3 shows the human attention distributions in different normalized vertical positions during the reading process, where we adopt the reading time of words to measure human attention and segment the reading process into four equal periods. It shows that in the first three periods, users usually read documents from top to bottom, regardless of whether there is an answer in the document, which is consistent with the findings in Figure 2.

Li et al. [18] classified the fixation transition behavior between lines into three types: *down*, *up* and *skip*. *Down* means the fixation moves from the currently fixated line to the next line. *Up* is the behavior that the fixation jumps to the lines above the current line. *Skip* represents that the fixation moves down with skipping several lines below. Figure 4(a) and 4(b) show the proportion of three types of transition behavior in with-answer and without-answer documents during four reading periods. In first three reading periods, the proportion of *down* behavior gradually decreases, while both *skip* and *up* increase in both with-answer and without-answer documents. This finding indicates that the user usually read sequentially in the early stages and gradually skim more text during the reading process.

Among 49 with-answer documents, 35 documents are annotated by users with multiple distinct answers. Therefore, in the sessions of these documents, besides the answer annotated by the user, there are other snippets which have been highlighted as answer evidence by other users. Therefore, for each session, we classify the text of its document into three categories: the answer annotated by the current user, the answers annotated by other users and non-answer text. Figure 4(c) shows the proportion of fixated text in 49 with-answer documents during four reading periods according to the reading time. We can see that although users have examined the snippets of candidate answers (i.e., the text finally annotated by themselves or other users) in the early reading process, they still continue reading and searching for answers.

In summary, we find an answer seeking stage in the early process of reading comprehension tasks with several clear reading behavior patterns. During this stage, users generally read the document from top to bottom and will continue reading until reaching the bottom of the document, although they may have examined some snippets of answer evidence. In addition, users tend to read the top of the document line by line and gradually skim with more *skip* and *up* transition behavior during a session.

4.2 Answer Verification

As shown in Figure 2 and Figure 3, after users have viewed the whole document sequentially in the answer seeking stage, there is still about 25% reading time left. As shown in Figure 3, in the last reading period, the overall attention distributions in with-answer documents and without-answer documents are significantly different at $p<0.01$ using two-tailed t-test test. For the without-answer documents, there are two distinct peaks in the attention distribution. The smaller peak at the top position states that users may return and reread the top of documents to verify their judgments that there is no answer in the documents, while the larger peak at the bottom position indicates that users may also leave reading directly when they reach the bottom and find no answer. For the with-answer documents, the attention distribution in the last period also has two peaks at the top and bottom positions, but most attention gathers at the top position, which reveals that in the last reading period of these documents, users are more likely to reread the documents. Then we look into the last reading period in Figure 4(c), where compared to the previous periods, users spend significantly longer time on reading snippets of candidate answers, including both the users' answers and others' answers, and pay less attention to non-answer texts. All these differences are statistically significant at $p<0.01$ using two-tailed t-test. In Figure 4(a) and 4(b), the proportions of three transition behavior are significantly different between with-answer and without-answer documents in the last reading period. In the with-answer documents, more *skip* and *up* behavior occurs, while in without-answer documents, only the proportion of *down* behavior increases compared to the previous periods.

From the analysis, we find that the user behavior patterns in the last reading period are much different from those in the first three periods. In the last reading period, users will reread more snippets of candidate answers with more *skip* and *up* transition behavior in the with-answer documents to make an answer verification before generating their final answers. In the without-answer documents, the user may also review the documents after reading the whole document but found no answer. Thus, we consider the last reading period as a distinct reading stage, i.e., answer verification stage.

4.3 Summary

Now we are able to answer **RQ1**. Figure 5 illustrates the two-stage reading behavior model during reading comprehension tasks. Given a question and a document, users first enter the answer seeking stage to read the whole document sequentially from top to bottom. If users cannot find an answer in the document, they may leave reading directly or review part of the document, while if users find several candidate answers in the document during the answer seeking stage, they will usually review these candidate answers to make a comparison and verification and then determine their final answers. Our two-stage reading behavior model supports the the multi-turn reasoning mechanism and the answer verification mechanism in the recent MRC models [8, 30, 31, 35, 37, 38].

5 READING BIAS

To address **RQ2**, we study four kinds of factors which may affect users' reading behavior: answer, position, lexical categories and matching signal.

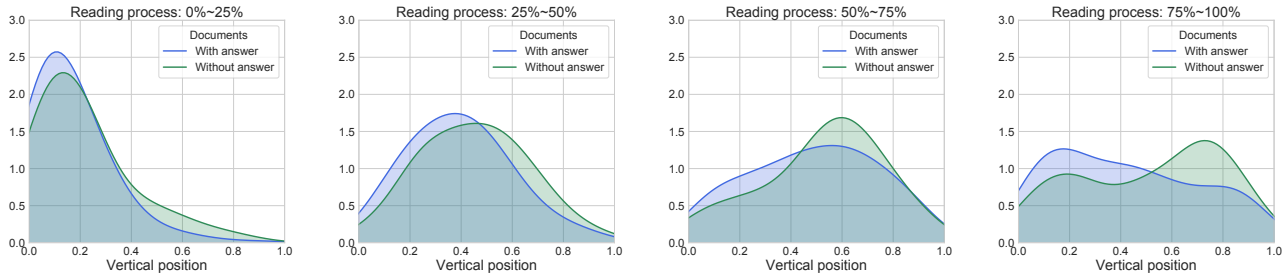
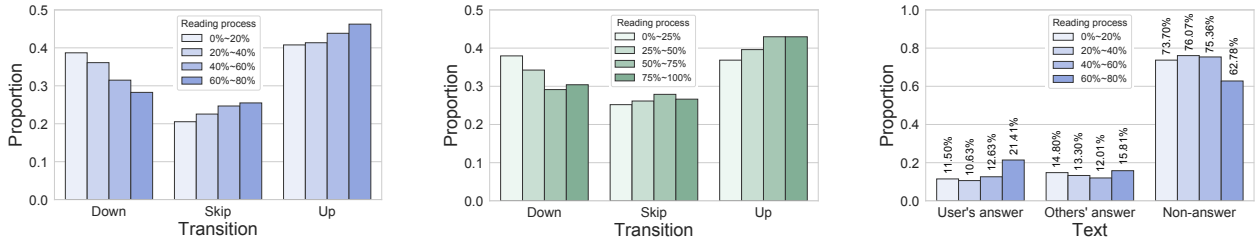


Figure 3: The overall attention distributions at different vertical positions of documents in four periods of the reading process.



(a) Transition in with-answer documents. (b) Transition in without-answer documents. (c) Fixated text in with-answer documents.

Figure 4: The proportion of transition behavior (4(a) and 4(a)) and fixated text (4(c)) in four periods of the reading process.

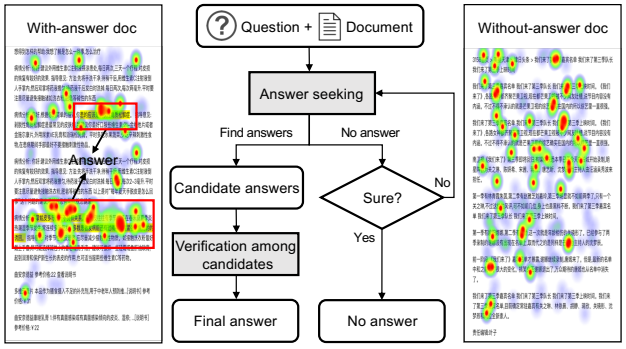


Figure 5: The two-stage reading behavior model with answer seeking (Stage 1) and answer verification (Stage 2).

5.1 Answer

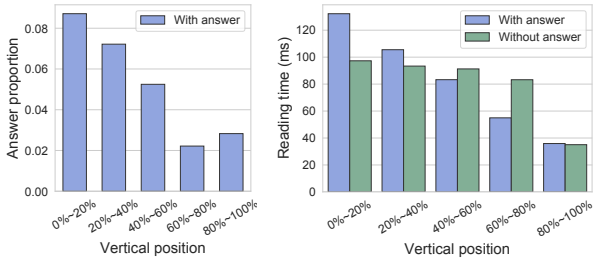
Table 2 shows the average fixation rate and reading time of words in documents. In with-answer documents, words are classified into two categories: answer or non-answer, according to whether they belong to the answer annotated by users. The results show that users tend to pay more attention to read answer word than non-answer word. Compared to words in without-answer documents, users spend less attention on non-answer word in with-answer documents. These findings can be observed in the fixation heat maps shown in Figure 5.

5.2 Position

Figure 6(a) shows the proportion of annotated answer words at five-level vertical positions in with-answer documents, showing that in the documents of our MRC tasks, the answer is more likely to appear in the top of the document than the bottom. Figure 6(b) shows the average reading time of words in documents with/without answers at five-level vertical positions. For the without-answer documents, we can see that at the first four vertical positions, the attention levels of users are similarly high with slight attenuation, while at the last vertical position, their attention declines sharply. This indicates that when users have not found any answer in documents, they

Table 2: The average fixation rate and reading time of words in documents with/without answer. All differences are statistically significant at $p < 0.01$.

Documents	With answer		Without answer
	Answer	Non-answer	
Fixation rate	0.938	0.263	0.293
Reading time (ms)	292.4	70.2	78.0



(a) Answer words. (b) Reading time.

Figure 6: 6(a) shows the proportion of answer words at five-level vertical positions and 6(b) shows the average reading time of words with respect to vertical positions.

Table 3: The average fixation rate and reading time of words with different IDF.

	Question				
Normalized IDF	0~0.2	0.2~0.4	0.4~0.6	0.6~0.8	0.8~1
Fixation rate	0.579	0.779	1.120	1.354	2.031
Reading time (ms)	215.2	276.4	401.9	555.6	846.6
	Document				
Normalized IDF	0~0.2	0.2~0.4	0.4~0.6	0.6~0.8	0.8~1
Fixation rate	0.151	0.218	0.314	0.360	0.476
Reading time (ms)	41.5	59.7	83.9	98.9	134.2

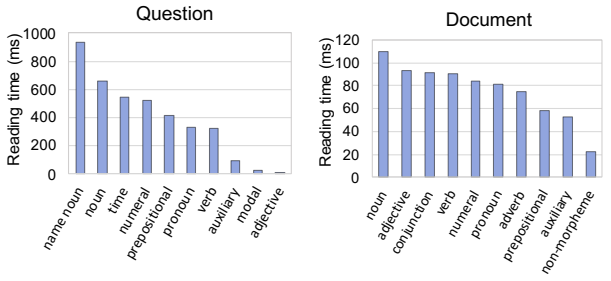


Figure 7: The average reading time of words with the first ten most frequent POS tags in questions and documents.

will concentrate in the reading process and their attention is hardly affected by the factor of vertical position except when they are reading the bottom of documents. This phenomenon can be observed in the heat map of the without-answer case shown in Figure 5. For the with-answer documents, we can see the distribution of reading time at five-level vertical positions is similar to the distribution of answer words in Figure 6(a). Their Pearson correlation coefficient is 0.966, indicating that human attention is biased by both vertical positions and answer locations. It is noteworthy that the position bias is also obvious at the end of the with-answer documents.

5.3 Lexical Features

Existing works show human attention during the reading process is affected by lexical features, such as word frequency [27] and Part-of-Speech tags [2]. Thus, we investigate the effects of these two features to human attention in reading comprehension tasks. Table 3 shows the average fixation rate and reading time of words with different IDF. We estimate the IDF of words based on the whole DuReader dataset [14] and normalize the IDF values of words within a document into 0 to 1. The Larger the IDF of a word is, the lower frequency the word is. From the results, we can see that low-frequency words attract more attention than high-frequency words when users are reading both questions and documents.

Figure 7 shows the average word reading time of the first ten most frequent POS tags in questions and documents. It shows that users tend to pay more attention to nouns during reading both questions and documents, because nouns usually contain useful and important information [7]. For those relatively less meaningful POS tags of words, such as modal, auxiliary and non-morpheme, users usually don't pay much attention on reading them.

Table 4: The average fixation rate and reading time of words with different cosine similarity between the question.

Cosine similarity	-0.25~0	0~0.25	0.25~0.5	0.5~0.75	0.75~1
Fixation rate	0.132	0.257	0.445	0.515	0.558
Reading time (ms)	36.2	68.8	123.5	143.7	157.8

Table 5: The results of ANOVA analysis on the effects of factors on human attention at the word level. The sample size is 34,560.

Factor	F	p	η^2 ($\times 10^{-3}$)
Answer	565.38	<0.001	**103.16**
IDF	179.77	<0.001	53.30
Position	201.36	<0.001	36.74
POS tag	63.92	<0.001	13.12
Cosine similarity	26.27	<0.001	6.59

5.4 Matching

Advanced neural MRC models capture the matching signals between the question and the document based on word embeddings. Thus, we would like to investigate the effect of the semantic matching signal on human attention. We train 300-dimension word embeddings on the whole Dureader dataset using Skip-gram [22]. The representation vector of a question $\mathbb{V}_Q$ is calculated as follows:

$$\mathbb{V}_Q = \sum_{q \in Q} IDF_q \times \mathbb{V}_q \tag{1}$$

, where Q is the question consisting of several question terms and $\mathbb{V}_q$ is the embedding vector of query term q. The IDFs of question terms are used to measure their importance in the whole representation of the question. Then we calculate the cosine similarity between vectors of the question and each document word as the matching signal. Table 4 shows the average word reading time of words with different cosine similarity, showing that words that are more semantically similar to the question usually attract more human attention.

5.5 ANOVA Analysis

All these factors can be classified into two categories: question-dependent and question-independent. The first category includes *answer* and *cosine similarity*, while the second one contains *position*, *IDF* and *POS tag*. To better answer **RQ2**, we conduct an ANOVA analysis to investigate the effects of five factors on human attention (reading time of words) during reading comprehension tasks. Table 5 reports the F-value, p-value and η^2 (effect size) of all factors. Although the p-values of all factors indicate statistical significance, the effect sizes of *POS tag* and *cosine similarity* are much smaller than other three factors. Among all factors, *answer* as a question-dependent factor is the most influential factor on human attention, followed by *IDF*, a question-independent factor. These findings also show the possibility of predicting human attention through the features of these factors and the feasibility of better locating answers in the document with the help of human attention.

6 ANSWER PREDICTION

To investigate **RQ3**, we leveraging human attention into MRC tasks to study whether it can help improve the performance. Since our aim is not to beat the state-of-the-art MRC model, we simplify our experiment settings by considering sentence as the minimum unit of an answer, and conduct the answer prediction experiment at the sentence level. The following experiment can be divided into two steps. First, we use the valid features in the two-stage reading behavior model to predict the human attention signals of sentences within a document. Second, we feed the predicted attention as a feature into answer sentence retrieval models to compare with baseline models.

6.1 Attention Prediction

We use five categories of features in the attention prediction task: *position*, *linguistic* and *matching*, *mouse* and *context*.

Position. Besides the position offset of the sentence, we calculate the normalized vertical and horizontal positions for each word of this sentence and use the max/min/mean/std values of word-level positions as the features of the sentence.

Linguistic. TF-IDF and word surprisal mean max/min/mean/std values of frequency in the document, IDF and surprisal of words in the sentence. For Part-of-Speech, we first obtain the first ten most frequent POS tags in all 60 documents, including noun, adjective, verb and etc. Then we use the occurrence ratios of these POS tags in the sentence as Part-of-Speech features.

Matching. We capture the matching signals between sentence and question from two aspects: exact matching and semantic matching. Question exact matching represents the occurrence number of query terms in the sentence, while the max/min/mean/std values of cosine similarities between sentence words and the question (See details in Section 5.4) is calculated for semantic matching.

Mouse. During the reading process, mouse movements are easy to obtain, which can be considered as a kind of feedback from users during reading comprehension tasks. We use the max/min/mean/std values of words' hover duration as extra human behavior features in attention prediction.

Context. Yang et al. [40] proposed to take advantage of contextual information in answer sentence retrieval by adding the features of the previous and next sentences into the features of the current sentence, which enhanced the model performances. In this work, we follow them and employ this approach in our models.

In the attention prediction task, we use the average reading time of words in a sentence as the label. The labels of sentences range from 0 to a few seconds. Therefore, we regard this task as a regression problem using Gradient Boosting Regression Tree (GBRT) [13] and an RNN-based neural regression model. We implement the RNN model with Gated recurrent unit (GRU) [5] and use the sequence of sentence hand-crafted features in a document as the input. Next, the outputs of GRU, which denote the learned representation vectors of sentences updated by contextual information, are fed into a multilayer perceptron (MLP) to predict the attention on each sentence. The loss function of the RNN model is Mean Square Error (MSE). The hidden size of GRU is 200 and the MLP have two hidden layers whose sizes are both 200. We set the dropout rate of GRU as 0.2 and use ReLU as the activation function in the MLP. A 5-fold cross validation at the document level, where the sentences of the same document belong to the same fold, is conducted to predict the attention of each fold and Pearson's Correlation Coefficient (PCC) is adopted to evaluate the performance. We use the two-tailed t-test to examine the significance of the differences between the predictions of different methods.

Table 6: The average 5-fold performances of attention prediction models. All the differences are statistical significance at $p < 0.01$ level using two-tailed t-test.

Feature Category	GBDT	RNN
Position	0.3115	0.4761
Linguistics	0.2297	0.3474
Matching	0.2111	0.3238
Mouse	0.2575	0.3969
All without context	0.4407	0.6316
All	**0.4880**	**0.6499**

Table 6 shows the average 5-fold performances of attention prediction models based on different categories of features. We can see that among four feature categories, *position* is the most effective one, which significantly outperforms the other three categories, followed by *mouse* features. *Matching* is the worst one on both GBDT and the RNN model. When combining the four categories of features together, both models achieve statistically significant improvement. Our results also show that when taking advantage of the contextual information, the performances of both GBDT and RNN model get significantly better, but the improvement of RNN is much smaller than that of GBDT. We consider that this is because the RNN model can automatically capture the contextual information by GRU without adding context features manually.

In this experiment, we show the effectiveness of the four categories of features in predicting human attention, which is consistent with our findings in the previous analysis. Finally, we obtain the best attention estimations from the RNN-based model.

6.2 Answer Sentence Retrieval

In the second experiment, we try to retrieve the answer sentences in the documents. This experiment is conducted on the 340 valid sessions whose answers are not "None". We classify the features used in this task into four categories: *learning to rank (LTR), matching*, *context* and *attention*. *Matching* and *context* are similar to those in attention prediction (Section 6.1), so here we highlight the other two kinds of features:

Learning to rank. We extract the same 8-dimension learning-to-rank features of each sentence as Qin and Liu [25], including sentence length (the number of words in the sentence), the average TF, IDF and TF×IDF values of query terms in the sentence, scores of BM25 and three language models with the question.

Attention. We use the average reading time of words in a sentence as the attention feature because this feature is the best one among all the attention features which we have tried, such as the max/min/mean/std values of words' reading time in a sentence and the total reading time of a sentence. We also apply two kinds of attention: real human attention collected in our user study (Section 3) and predicted attention by models (Section 6.1). For the latter one, we choose the attention predicted by the best-performing RNN model with all categories of features (Table 6).

Table 7: The Performances of answer sentence retrieval models. The *real* attention denotes the feature extracted from human attention collected in our user study. The *predicted* attention is estimated by the best model in the attention prediction task. */ indicate statistical significance at p < 0.05/0.01 level compared to the same model without attention.**

Model	Attention	BLEU-4	Rouge-L	nDCG@1	nDCG@3	nDCG@5
BM25	-	16.19	27.71	0.183	0.293	0.351
MART	-	19.70	30.06	0.321	0.376	0.407
RankBoost		19.70	31.61	0.324	0.382	0.459
LambdaMart		20.50	32.02	0.358	0.384	0.426
RNN		18.43	29.68	0.364	0.395	0.444
MART	Predicted	23.79	**38.75****	0.347	0.414	0.442
RankBoost		**25.69***	37.78*	0.350	0.407	0.452
LambdaMart		23.84	35.98	0.376	0.425	0.452
RNN		23.51**	37.03**	**0.381**	**0.441**	**0.486**
MART	Real	30.60**	42.55**	0.571**	0.645**	0.682**
RankBoost		**33.79****	42.35**	0.673**	0.723**	**0.738****
LambdaMart		32.27**	42.72**	**0.694****	**0.724****	0.736**
RNN		33.69**	**43.60****	0.619**	0.707**	0.718**

In this task, the label is the annotation rate of sentences, which we scale into three grades: 0←0, 1←(0, 0.5], 2←(0.5, 1], which account for 83.4%, 13.2% and 7.4% respectively. For the ranking problem, we use five retrieval model: BM25, MART(i.e., GBDT), RankBoost[12], LambdaMart[39] and a RNN-based neural ranking model which shares the same framework as the one in the attention prediction task. The partition of 5-fold cross validation here is also the same as that in the attention prediction task. To evaluate the model performance in the MRC task, we choose the BLEU-4[24] and Rouge-L[19], where all the user answers (i.e., the answers generated by the participants) in the task are used as the reference answers and the top k retrieved sentences are concatenated according to the positional order to serve as the predicted answer. Since the number of average annotated sentences in sessions is 2.1, we set k as 2. For the ranking performance, because the average number of sentences whose labels are greater than zero in a document is 4.4, we report nDCG at position 1, 3 and 5.

Table 7 shows the performances of answer sentence retrieval models. We first look into whether real human attention can help improve the model performance in the MRC task. When compared to BM25 and those baseline models without attention, the models with real human attention have statistically significant improvements on both MRC and ranking metrics, which is consistent with our previous findings in Section 5.1. Further, we use the predicted human attention as an alternative to real human attention and find that the models with predicted attention still perform better on all the MRC and ranking metrics than baseline models without attention. With predicted attention, the RNN-based model achieves statistically significant improvement on both MRC metrics compared to itself without attention.

To address **RQ3**, we use two kinds of human attention as an extra feature in answer retrieval models, which improve the performances in all MRC and ranking metrics. Although predicted human attention is less effective than real human attention, it also can lead models to a better performance in MRC tasks. Our experiment shows that learning from features of human behavior can help improve the performance of MRC tasks.

7 CONCLUSION

In this study, we thoroughly investigate human behavior patterns during reading comprehension tasks. We conduct a lab-based user study to collect human behavior data in reading comprehension tasks. By analyzing the collected eye movements and answer annotations, we propose a two-stage reading behavior model, where human tends to first seek answers in the document and then make verification among candidate answers. We study several question-dependent and question-independent factors that may affect human attention during reading comprehension tasks, such as answer location and the cosine similarity with the question, vertical position, IDF and POS tags of words. Our analysis shows that the attention distributions of humans strongly correlate with the answer location and are biased by other factors. In the answer prediction task, we show the effectiveness of real human attention. Then we utilize features of human attention factors to predict human attention and incorporate the predicted attention into answer retrieval models as extra features, which leads to better performances on all MRC and ranking metrics in comparison with baseline models without human attention features.

The two-stage reading behavior model can help to better understand how human reads and seeks answers during reading comprehension tasks and may inspire the MRC model to achieve better performance. In the future, we would like to study two aspects. The first one is how the candidate answers that have been examined influence human's subsequent reading behavior during the reading comprehension task. The second future work is to compare the human behavior in a lab-based user study with that in the real-life reading comprehension scene such as finding answers for a certain question in the search engine. The two aspects can help to better understand the reading process of human and improve the search engine and QA systems.

8 ACKNOWLEDGEMENTS

This work is supported by Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011) and the National Key Research and Development Program of China (2018YFC0831700).

REFERENCES

[1] Jeanette Altarriba, Judith F Kroll, Alexandra Sholl, and Keith Rayner. 1996. The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. *Memory & Cognition* 24, 4 (1996), 477–492.

[2] Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

[3] Klinton Bicknell and Roger Levy. 2010. A rational model of eye movement control in reading. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*.

[4] Georg Buscher, Andreas Dengel, Ralf Biedert, and Ludger V Elst. 2012. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 2 (2012), 9.

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[6] Uschi Cop, Denis Drieghe, and Wouter Duyck. 2015. Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. *PloS one* 10, 8 (2015), e0134008.

[7] Robert G Crowder and Richard K Wagner. 1992. *The psychology of reading: An introduction*. Oxford University Press.

[8] Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2016. Gated-attention readers for text comprehension. *arXiv preprint arXiv:1606.01549* (2016).

[9] Ralf Engbert, André Longtin, and Reinhold Kliegl. 2002. A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision research* 42, 5 (2002), 621–636.

[10] Ralf Engbert, Antje Nuthmann, Eike M Richter, and Reinhold Kliegl. 2005. SWIFT: A dynamical model of saccade generation during reading. *Psychological review* 112, 4 (2005), 777.

[11] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[12] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *Journal of machine learning research* 4, Nov (2003), 933–969.

[13] Jerome H Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[14] Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2017. Dureader: A Chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073* (2017).

[15] Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review* 87, 4 (1980), 329.

[16] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.

[17] Jiajia Li, Grace Ngai, Hong Va Leong, and Stephen CF Chan. 2016. Your eye tells how well you comprehend. In *Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual*.

[18] Xiangsheng Li, Yiqun Liu, Jiaxin Mao, Zexue He, Min Zhang, and Shaoping Ma. [n. d.]. Understanding reading attention distribution during relevance judgement. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*.

[19] Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*.

[20] Yiqun Liu, Chao Wang, Ke Zhou, Jianyun Nie, Min Zhang, and Shaoping Ma. 2014. From skimming to reading: A two-stage examination model for web search. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*.

[21] Yiqun Liu, Xiaohui Xie, Chao Wang, Jian-Yun Nie, Min Zhang, and Shaoping Ma. 2017. Time-aware click model. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 16.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[23] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*.

[25] Tao Qin and Tie-Yan Liu. 2013. Introducing LETOR 4.0 datasets. *arXiv preprint arXiv:1306.2597* (2013).

[26] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[27] Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124, 3 (1998), 372.

[28] Erik D Reichle, Alexander Pollatsek, Donald L Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological review* 105, 1 (1998), 125.

[29] Erik D Reichle, Keith Rayner, and Alexander Pollatsek. 2003. The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences* 26, 4 (2003), 445–476.

[30] Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. Reasonet: Learning to stop reading in machine comprehension. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[31] Alessandro Sordoni, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2016. Iterative alternating neural attention for machine reading. *arXiv preprint arXiv:1606.02245* (2016).

[32] Sara Stymne, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull, and Martin Wester. 2012. Eye tracking as a tool for machine translation error analysis. In *LREC*.

[33] Chuanqi Tan, Furu Wei, Nan Yang, Bowen Du, Weifeng Lv, and Ming Zhou. 2017. S-net: From answer extraction to answer generation for machine reading comprehension. *arXiv preprint arXiv:1706.04815* (2017).

[34] Takenobu Tokunaga, Hitoshi Nishikawa, and Tomoya Iwakura. 2017. An eye-tracking study of named entity annotation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP*.

[35] Adam Trischler, Zheng Ye, Xingdi Yuan, Philip Bachman, Alessandro Sordoni, and Kaheer Suleman. 2016. Natural language comprehension with the EpiReader. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

[36] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

[37] Yizhong Wang, Kai Liu, Jing Liu, Wei He, Yajuan Lyu, Hua Wu, Sujian Li, and Haifeng Wang. 2018. Multi-passage machine reading comprehension with cross-passage answer verification. *arXiv preprint arXiv:1805.02220* (2018).

[38] Dirk Weissenborn. 2016. Separating answers from queries for neural reading comprehension. *arXiv preprint arXiv:1607.03316* (2016).

[39] Qiang Wu, Christopher JC Burges, Krysta M Svore, and Jianfeng Gao. 2010. Adapting boosting for information retrieval measures. *Information Retrieval* 13, 3 (2010), 254–270.

[40] Liu Yang, Qingyao Ai, Damiano Spina, Ruey-Cheng Chen, Liang Pang, W Bruce Croft, Jiafeng Guo, and Falk Scholer. 2016. Beyond factoid QA: Effective methods for non-factoid answer sentence retrieval. In *European Conference on Information Retrieval*.

[41] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining local convolution with global self-Attention for reading comprehension. *arXiv preprint arXiv:1804.09541* (2018).