Do People and Neural Nets Pay Attention to the Same Words? Studying Eye-tracking Data for Non-factoid QA Evaluation

Valeriia Bolotova
RMIT University
lurunchik@gmail.com

Vladislav Blinov
Ural Federal University
vladislav.blinov@urfu.ru

Yukun Zheng
Tsinghua University
zhengyk13@gmail.com

W. Bruce Croft
University of Massachusetts Amherst
croft@cs.umass.edu

Falk Scholer
RMIT University
falk.scholer@rmit.edu.au

Mark Sanderson
RMIT University
mark.sanderson@rmit.edu.au

ABSTRACT

We investigated how users evaluate passage-length answers for non-factoid questions. We conduct a study where answers were presented to users, sometimes shown with automatic word highlighting. Users were tasked with evaluating answer quality, correctness, completeness, and conciseness. Words in the answer were also annotated, both explicitly through user mark up and implicitly through user gaze data obtained from eye-tracking. Our results show that the correctness of an answer strongly depends on its completeness, conciseness is less important.

Analysis of the annotated words showed correct and incorrect answers were assessed differently. Automatic highlighting helped users to evaluate answers quicker while maintaining accuracy, particularly when highlighting was similar to annotation. We fine-tuned a BERT model on a non-factoid QA task to examine if the model attends to words similar to those annotated. Similarity was found, consequently, we propose a method to exploit the BERT attention map to generate suggestions that simulate eye gaze during user evaluation.

ACM Reference Format:
Valeriia Bolotova, Vladislav Blinov, Yukun Zheng, W. Bruce Croft, Falk Scholer, and Mark Sanderson. 2020. Do People and Neural Nets Pay Attention to the Same Words? Studying Eye-tracking Data for Non-factoid QA Evaluation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340531.3412043

1 INTRODUCTION

Search Engine Result Pages (SERP) commonly display an answer as well as a list of retrieved documents. Much research has been conducted on the formation of such an answer and also on ideal SERP layout. Research on whether different answer presentation styles allow users to obtain information more quickly or accurately is more limited and mostly considers answers on factoid questions which could be answered with relatively short snippets. Few works have examined non-factoid questions, which often require a passage-level answer. Do users look for the same words or facts when they evaluate the answer to such a question? Does too much or too little detail in an answer impact answer quality?

One exception is Qu et al.'s study, which employed crowdsourcing to determine if workers could identify the correctness of answers [25]. The researchers examined if identification could be improved by automatic highlighting of specific answer words. The results were inconclusive as Qu et al. did not consider what aspects of an answer contribute to overall quality, or whether one answer presentation style was more effective than another.

We seek to understand how users determine the correctness of an answer by considering fine-grained evaluation criteria and employing an eye-tracker. We compare different tracking metrics on answers to non-factoid questions with or without highlighting. Users were also asked to mark up important words in the answer and provide feedback. We also analyse the attention of a BERT [10] model fine-tuned on a non-factoid QA evaluation task. Although many works have analysed the attention mechanism of different Transformer models [6, 19, 31], we focus on comparing the model's attention with words annotated in the answer based on explicit and implicit input. We chose the BERT model for our experiments due to its success in the context of non-factoid question answering [22].

Our work investigates three research questions:

(1) How do people understand whether an answer for a non-factoid question is correct? What features of an answer determine overall answer quality, and is the evaluation process similar across users and answers of different quality?
(2) Can automatic word highlighting improve the speed and accuracy of users when determining answer correctness?
(3) Does the attention map of a Transformer machine learning model assign weights to words in a similar way to user annotated words? Can we use the weights to highlight important words in passage-level answers to non-factoid questions?

The contributions of our paper are as follows.

(1) We demonstrate that user knowledge of an answer to a non-factoid question influences perceptions of accuracy. The completeness of an answer is as important as correctness. According to user annotations and eye-tracking data, there is agreement on the parts of an answer that contribute to correctness. Also, it is simpler for users to identify incorrectness of an answer than conclude that it is correct.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412043

(2) We show that word highlighting in an answer for a non-factoid question helps users to evaluate an answer faster while maintaining the same accuracy.
(3) We analyze BERT model attention in a new way, directly comparing the attention of a BERT model with human attention (both explicit annotations and implicit eye-tracking data). We propose an algorithm[2] that constructs word highlighting from BERT model weights, and show that the new highlighting has higher similarity with user annotations compared to a baseline method.

2 RELATED WORK

There has been extensive research on examining the generation of answers to factoid questions, ranking of sentences, modeling of community question answering sites, as well as identifying passages that answer complex questions [8, 15, 18, 24, 30, 40, 41]. Here, however, we consider less common research on the display of answers and user reactions to them. We also examine the utilization of eye-tracking to understand how users interact with passage-level answers to non-factoid questions.

2.1 Constructing and Displaying Answers

Before answers, there were snippets: fragments of a retrieved document that matched a user's query. The creation and value of snippets was demonstrated over twenty years ago [32]. Later, the impact of different snippet designs on user clickthrough was examined [7]. Results indicated the importance of users seeing all possible query terms in the snippet so that they could judge the terms' relationships to the content of a page. Eye-tracking was subsequently employed to gain clues on how snippets might be better constructed [3].

Building on Clarke et al. [7]'s work, Iofciu et al. [14] examined two approaches to highlighting words in a snippet: query words in bold, additional words in color. The work focused on queries with ambiguous intent: words that identified intent were highlighted in the snippet. Iofciu et al. used two approaches to word choice: manual and automatic. Lab-based experiments with manual highlighting found users were slower and less accurate in their clicking. Automated testing employed analysis of a query log to better identify ambiguous queries and the logs plus Wikipedia disambiguation pages provided information on words to highlight. With this form of highlighting, users were found to click accurately and faster when highlighting was present.

Qu et al. [25] considered if highlighting a suite of words in an answer would allow users to identify good or bad answers more accurately. Utilizing crowd-source workers, the researchers found that highlighted words appeared to influence the workers decision, but results were not conclusive.

2.2 Eye-tracking and User Search Interaction

We detail two types of studies: those concerned with how users interact with search results and those that use eye-tracking to understand the design of snippets.

Granka et al. [11] showed that a user's gaze fixated mainly on the first and second ranks of a SERP. Lorigo et al. [21] considered the impact of re-ranking results on a user's gaze. Using eye-tracking to understand the way that users scanned a SERP allowed refinements to rank learning algorithms, which improved effectiveness [16]. Eye-tracking also informed how users interact with SERPs [13, 37].

Cutrell and Guan [9] used eye-tracking to explore the impact of the length of snippets on the speed and accuracy with which users could complete tasks. They found that longer snippets helped with informational tasks, but hindered navigational [4]. Savenkov et al. [27] showed that highlighting of terms "*help users find the answer faster and draw their attention to results in the lower part of SERP*". Lagun et al. [20] investigated the relationship between eye gaze and the browser viewport in mobile search for answer-like results from factoid questions. Some studies researched how direct answers to frequent search queries, such as weather and news, influence user behavior and how they contribute to satisfaction [5, 38].

To the best of our knowledge, there is no research on the use of eye-tracking to understand how users interact with non-factoid answers either presented on their own or in a SERP.

Table 1: Three examples of questions used in the study.

TaskID	Question
14579170	How to get a speeding ticket dismissed?
17172970	Why is my MRSA Staph infection continuing to relapse?
17636401	How to check for dead pixels when buying psp?

3 USER STUDY

We conducted a lab-based user study observing interactions with a question answering system with/without automatic highlighting of words. The study was ethically approved by RMIT University.

3.1 Tasks and Users

We randomly sampled 40 questions from the existing nfL6 dataset[1] [25] consisting of content from Yahoo!Answers. Table 1 shows three examples of questions used in our study. Each question was paired either with an answer selected as "best" by the (Yahoo!) asker of the question or it was paired with an incorrect answer. This was obtained by ranking answers using BM25 scoring [29], which were re-ranked using a BLSTM model [8]. The distributions of answer lengths is similar for correct and incorrect answers, with a mean (and standard deviation) of 344.0 (152.3) and 419.7 (197.4), respectively. We automatically highlighted words in the answers based on a past approach [25], highlighting the five words with the highest *TF-IDF* weights and all capitalized words (excluding those starting sentences). The final QA dataset is available here[2].

We recruited 32 candidates. After checking visual acuity (using a Tobii Pro X2-60 eye-tracker) to ensure that the collected eye movements would be of adequate quality, we chose 20 users: 15 females, 4 males, and 1 preferring not to disclose. All were either students or university staff and had good knowledge of English, 7 participants in age-group (18-24), 5 in (25-30), 2 in (31-35), 4 in (36-40) and 2 participants over 40. They spent around one and a half hours on average on the tasks. After successful completion of 40 tasks, each user received a $50 (USD) gift voucher as compensation.

[1] https://ciir.cs.umass.edu/downloads/nfL6/
[2] https://github.com/Lurunchik/non-factoid-answer-highlighting

3.2 Procedure

After calibration with the eye-tracker, each user was guided through two training tasks. The users then completed the forty QA tasks, each of which involved five screens:

(1) **Question Reading:** A question was shown for the user to read. Pressing the space bar moved to the next screen (this interaction has less impact on eye-tracking data than a clickable button).
(2) **Pre-task Questionnaire:** The user needed to complete a pre-task questionnaire about their *interest*, *familiarity*, and *perceived difficulty* of the question on a scale of 1-5 (low-high). They could re-read the question up to twice if desired.
(3) **Answer Reading:** An answer was displayed for the user to read. Answers consisted of four sentences, no scrolling was needed. Pressing the space bar moved to the next screen.
(4) **Post-task Questionnaire:** The user was asked to rate the *overall quality* of the given answer, on a scale of 0-2, as well as its *correctness*, *completeness*, and *conciseness* on a scale of 1-5 (low-high). The quality scale was enumerated as (0) An INCORRECT answer to the question; (1) A CORRECT but low-quality (*LQ*) answer (e.g. a partial answer); and (2) A CORRECT and high-quality (*HQ*) answer to the question. Additional user feedback could also be provided.
(5) **Word Annotation:** The answer was displayed again and the user was asked to markup positive and negative words, selecting them using a mouse. Users were instructed to mark as positive those words that convinced them the passage was correct; negative, those words that convinced users the answer was of low quality or incorrect. Users could markup any complete words, phrases, or sentences, but were advised at the start to markup in a fine-grained manner.

3.3 Experimental Setting

Users were randomly divided into two equal groups. The first assessed answers without automatic highlighting, the second was shown an answer with highlighted words as described earlier. For each question, its correct answer was shown to five out of ten users, and its incorrect answer was shown to the other five. In total, each answer was assessed by 10 users.

4 ANALYSIS OF ANSWER EVALUATION

Our user study resulted in data that includes user knowledge about a question, answer quality, explicit word markup by users, and implicit gaze on words based on eye-tracking. In this section, to study how people evaluate answers to non-factoid questions, we look at the quality rating, gaze metrics and differences between explicit and implicit annotations by users.

4.1 Overall Answer Quality

We first consider *overall answer quality* ratings (0-2) made by users. The average agreement for ratings between all pairs of users was moderate, with a Cohen's Kappa [1] value of 0.59.

To obtain a single quality label for each answer for subsequent analysis, we took the majority vote across the ten users. In total, 43 answers were rated as INCORRECT and 37 as CORRECT; of the latter group, 22 were rated as *LQ* and 15 as *HQ*. Recall that we constructed our data based on the initial nfL6 collection to include forty correct and forty incorrect answers; manual inspection showed the three discrepancies between our users' ratings and this dataset were due to labeling errors in nfL6 collection. These were answers rated as incorrect but marked as correct or vice-versa. *All subsequent analysis was based on this corrected grouping.*

Table 2: Accuracy of rating corresponding to ground truth.

Answer quality	User Settings	
	without highlighting	with highlighting
ALL	0.72	0.71
INCORRECT	0.81	0.77
CORRECT	0.62	0.64
LQ & HQ	0.58 & 0.67	0.54 & 0.8

Table 2 shows the accuracy of user ratings, defined as the ratio of answers labelled in accordance with the majority vote to the total number of rated answers. The accuracy is higher for incorrect answers and the most common mistake was when users rated a correct answer as an incorrect one (28 false positives versus 91 false negatives). For each answer that had at least one erroneous rating, we studied the difference between users who made a mistake and those who did not. We found the mean value of user perceived difficulty of questions was higher for users who made a mistake compared to those who didn't (2.8 vs 2.5). Mean values of interest and familiarity were lower for mistaken users (2.7 and 2.2, respectively) vs (3 and 2.4, respectively). Note, however these differences were not statistically significant. When considering user errors, there are two types: *false positives* and *false negatives*. We will study causes of these mistakes separately.

For the answers where *false negative* were made, there was a statistically significant increase in perceived difficulty of a question ($p = 0.04$) between users who mistakenly rated correct answers as incorrect and users who rated answers correctly. Therefore, *when a question is difficult for a user, they tend to rate a correct answer as incorrect probably because of a lack of information both in the answer and in the user's initial knowledge.*

Regarding *false positives*, there is a statistically significant decrease in the user's interest in the question between those who mistakenly rated incorrect answers as correct and those who correctly rated them ($p = 0.03$). This indicates that *users could be less attentive while rated the question they are not interested in.* Below there is an example of a QA pair with very high variance of answer quality ratings among users:

`Q: I have a dsl connection another house member is using it for a wireless connection how can I stop access 2 him?`

`A: Upgrade to dsl and tell them what you want to do or upgrade to cable and get yourself a router. With a dial-up connection, a router is a waste of time. It'll work but if you upgrade to dsl, chances are it won't cost you any more (or much more) than dial-up and they'll give you all the equipment you need. Plus you'll have a much better connection with dsl over dial-up.`

Here, three out of ten users rated the incorrect answer as *HQ*. All three users had an interest score of only one out of five. We hypothesise that they were misled by the high keyword overlap

Table 3: Mean answers quality ratings (scale from 1 to 5).

	correctness		completeness		conciseness	
	mean	kappa	mean	kappa	mean	kappa
INCORRECT	1.57	0.25	1.53	0.11	1.85	0.16
CORRECT *LQ*	3.2	0.01	2.82	-0.05	2.95	0.09
CORRECT *HQ*	4.25	0.03	3.93	0.1	3.54	0.14

between the question and the answer. Notably, two of the users only saw the plain text without highlighting.

4.2 Answer Quality Aspects

We also obtained ratings for three quality aspects: *correctness*, *completeness*, and *conciseness*. Table 3 provides detailed information on the mean values and weighted Cohen's Kappa agreement between users for each aspect. The ratings suggest that even high quality answers, which were chosen as the best on the Yahoo!Answers website, can be greatly improved. There is fair agreement on the rating of correctness for incorrect answers (the mean value is 1.57 with 0.25 Cohen's Kappa). However, there is almost no agreement for the rest.

Regarding the contribution of each answer quality aspect, both *correctness* and *completeness* have high correlation (Pearson, $p < 0.001$) with the overall quality rating, followed by moderate correlation for *conciseness*. The values for the three aspects were 0.85, 0.79, 0.63 measured across all answers. Separately, within only incorrect answers, the strongest correlation is correctness (0.72, 0.56 and 0.4). Within correct answers, correctness and completeness have comparable correlation with overall quality (0.79 and 0.77, 0.54).

There are few cases when correctness was assigned a high rating while completeness a low one, and we also found a high correlation between correctness and completeness (0.817) for all answers and 0.818 for only correct ones, $p < 0.001$. In comparison, correlation of correctness with conciseness was only moderate overall (0.63), and low within correct answers (0.47), $p < 0.001$. *This analysis indicates an important relationship: the answer to a non-factoid question is rated as correct only when it is also complete. For instance, the answer contains all parts of an explanation, or different opinions, examples, and so on. On the other hand, both completeness and conciseness have little meaning when an answer is incorrect and they were rated rather randomly by users.*

To understand what could influence user rating, we investigated the relationship between aspect score values and user perceptions of question difficulty, familiarity, and interest. Cases with low and high aspect score variance were analyzed separately, as sometimes the rating of an answer does not require extra knowledge, e.g. when an answer is obviously incorrect or correct. Below there is a correct Q&A pair with low score variance of answer quality (the variance is 0 overall and 0.1 for each aspects) but high variance (>2) of user familiarity with the question.

`Q: Why is ice less dense than water?`

`A: The molecules of water are closer together and constantly moving, whereas the molecules of ice are in a crystal lattice, meaning they're in a rigid formation. When water freezes, the molecules spread out a little more to form the crystal lattice. Since density is mass over volume, and ice has takes up more volume than water, the density of ice is lesser than that of water. Which makes ice float on water.`

Ideally, we want automatic QA-models to generate answers that contain enough information to be understood and assessed correctly regardless of a user's initial knowledge about the question. For answers with higher quality score variance, we only found low positive correlation (0.24 Pearson, $p < 0.05$) between correctness ratings and interest in the question. We found a low negative correlation (-0.21 Pearson, $p < 0.05$) between difficulty and conciseness. Thus, we did not observe any definitive strong influence there.

4.3 Gaze Analysis

During reading, people make a series of rapid eye-movements called *saccades*, while for some periods of time the eyes are relatively still, called *fixations*. It is during the fixations that a reader acquires information [39]. We therefore focus on fixations in our analysis of our users. The text reading process also involves fixations that go against the normal reading order of left to right and top to bottom for English text. In such cases, people return back to already seen parts of a text. Such fixations are called *regressions*.

From the raw eye-tracking data, we can obtain the duration of fixations and their positions on the screen. We analyse fixations that last for more than 60 milliseconds in all subsequent analysis, following common practice [42].

Eye movements can be influenced by many factors including reading ability [2], a person's prior knowledge about a question [17], and demographics [26]. However, it has been found that people focus their attention on words that are relevant to their question [27]. We studied word-level fixations, and for each word, we identified all fixations whose coordinates fell into a word-sized bounding box. The following gaze measures were used to study how our users interacted with answers of different quality for a non-factoid question:

- *Total view seconds*, the total time spent looking at a screen with an answer.
- *Mean fixation duration*, the average length of fixations, per user and per answer.
- *Mean regression duration*, the average length of fixations which were regressions to a word, per user and per answer.
- *Mean word fixations*, the average number of fixations on a word, per user and per answer.
- *Mean word regressions*, the average number of word regressions in the answer, per user and per answer.

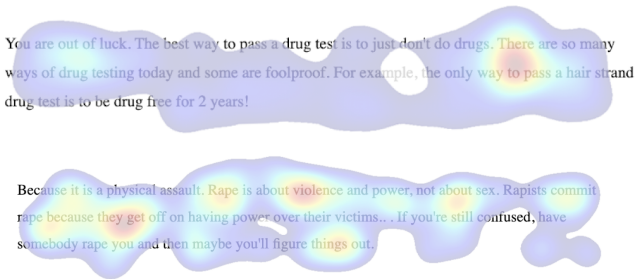

Figure 1: Correct/incorrect (top/bottom) answer heatmap.

Table 4: Gaze metrics for different answer quality levels.

Ans. quality	Metric: Total view seconds	Mean fixation duration	Mean regression duration	Mean word fixations	Mean word regressions
			among all users		
ALL	17.62	174.89	174.8	61.95	24.28
INCORRECT	16.89*	173.62	172.1*	60.22	22.83*
CORRECT	18.48*	176.37	177.93*	63.96	25.97*
LQ & HQ	17.39 & 20.08	176.5 & 176.17	179.75 & 175.256	60.61 & 68.87	25.31 & 26.93
			without guiding highlighting		
ALL	**18.66**‡	**181.8**‡‡	**181.14**‡‡	**63.12**	**25.44**
INCORRECT	18.18‡	180.22‡‡	179.29‡‡	61.81	24.46
CORRECT	19.21	183.63‡‡	183.25‡‡	64.63	26.58
LQ & HQ	16.98 & 22.47‡	182.26‡‡ & 185.68‡‡	183.34‡ & 183.12‡‡	59.02 & 72.87	24.11 & 30.2
			with guiding highlighting		
ALL	**16.62**‡	**168**‡‡	**168.59**‡‡	**60.87**	**23.21**
INCORRECT	15.59*‡	167.01‡‡	165.11*‡‡	58.63	21.2*
CORRECT	17.81*	169.15‡‡	172.67*‡‡	63.48	25.54*
LQ & HQ	17.8 & 17.83‡	170.74‡‡ & 166.8‡‡	176.21‡ & 167.4‡‡	62.2 & 65.36	26.52 & 24.09
			without guiding highlighting (users with 100% labeling accuracy)		
ALL	**18.31**	**181.15**††	**180.80**††	**63.91**	**26.19**
INCORRECT	18.92	179.61††	178.29††	64.05	26.91
CORRECT	17.61	183.5††	184.54††	63.76	25.37
LQ & HQ	16.04 & 19.92	183.37† & 183.69††	187.88†† & 180.33†	57.73 & 72.61	22.43 & 29.69
			with guiding highlighting (users with 100% labeling accuracy)		
ALL	**16.92**	**167.59**††	**165.93**††	**61.88**	**24.1**
INCORRECT	16.13*	165.5*††	162.29*††	60.71	23.167*
CORRECT	17.82*	170.49*††	170.98*††	63.22	25.17*
LQ & HQ	18.47 & 16.92	171.65† & 169.35††	173.64†† & 168.36†	62.81 & 63.79	27.26 & 22.24

* *significant difference between correct/incorrect answer groups;* ‡ *difference between users shown/not shown highlighting;*
† *difference between 100% accurate users shown/not shown highlighting. One symbol* $p < 0.05$*, two symbols* $p < 0.001$.

All results for these gaze metrics are shown in Table 4. The values for total view seconds, mean fixation, and regression duration are all lower for incorrect answers, which demonstrates that *users spend more time and effort to understand that an answer is fully correct than that it is missing some information or is incorrect.* Word fixations and regressions, which have been shown in related work to be indicators of relevance [12], are lower for wrong answers. We tested the statistical significance of differences between metric values for correct and incorrect answers using Student's t-test. Significant cases were marked in Table 4. Results were statistically significantly different for total view seconds, regression duration, and word regression counts. This finding is consistent with results from Qu et al. [25] that *people interact with good and bad answers differently, and rate incorrect answers with less effort.*

Typically, incorrect answers also have people gaze at more areas of interest with lower average fixation duration compared to correct answers. An illustrative example of gaze heatmaps for a who who was not shown highlighted words is shown in Figure 1.

4.4 Words Annotated While Answer Rating

For factoid questions, which typically have short answers, a set of important words for understanding if an answer is correct may comprise of most or all of the answer text. On the other hand, it is not obvious which parts of a passage-level answer for non-factoid question help people to evaluate the answer correctness. Our study users were asked to freely annotate positive and negative words and sections of text in the answer. While they were rating an answer, their gaze was recorded using the eye-tracker. In this subsection we investigate the level of agreement between users regarding the sets of words that were annotated for answer evaluation for both explicitly marked up words, and words that were gazed at. We also studied whether a user's explicit annotations were similar to those words as tracked from their gaze.

4.4.1 Explicit word annotation. Users were able to freely annotate positive and negative words which influenced their decision about the correctness of an answer. Only a tiny proportion of words (0.02%) was annotated in opposite annotation from the final answer rating. Correct answers were annotated with positive words, incorrect answers were annotated with negative words. Most opposite annotations were positive words annotated in incorrect answers. This suggests that correct answers that were rated *LQ* were generally missing information, rather than presenting incorrect facts. The relative length of annotations words in correct answers was statistically significantly higher than the length in incorrect ones. This suggests that *people struggle to identify the wrong parts of an incorrect answer to a non-factoid question.* There was a statistically significant positive correlation between answer quality aspect rating and the relative lengths of annotations: 0.33, 0.29 and 0.46 for correctness, completeness and conciseness respectively (Pearson, $p < 0.05$).

Table 5: Overlap between explicit important annotation.

Answer quality	User settings: all	without highlighting	with highlighting	between 2 settings
ALL	0.66	$0.68^{\dagger}$	$0.7^{\dagger}$	0.64
INCORRECT	0.65^{*}	$0.66^{\dagger *}$	$0.69^{\dagger}$	0.62^{*}
CORRECT	0.68^{*}	0.71^{*}	0.7	0.66^{*}
LQ & HQ	0.67 & 0.7	0.69 & 0.73	0.71 & 0.7	0.64 & 0.68

** significant difference between correct/incorrect answer groups;*
† between users shown/not shown automatic highlighting ($p < 0.05$).

To calculate the agreement between annotated spans of text in passages, we followed the approach of Qu et al. [25] and used the overlap coefficient [35]:

$$Overlap(H_1, H_2) = \frac{|H_1 \cap H_2|}{min(|H_1|, |H_2|)}$$

where H_1 and H_2 are unique words from the annotations of two users. This metric allows us to compare annotations of different lengths as some users annotated distinct words and others whole sentences. Following Qu et al. [25], we excluded stopwords. We calculated overlap only between positive words for correct answers, and negative words for incorrect answers because of the almost complete lack of opposite annotations.

Table 5 shows the mean overlap scores between user pairs, broken down by different user setting groups (columns) and different answer quality ratings (rows). Overall, the agreement for correct answers was higher, which was consistent with Qu et al. [25] and means that *users had good agreement between each other on the words that were important for the identification of a correct answer.*

4.4.2 Implicit word annotation. Using the eye-tracking data, we can calculate the agreement between words that user annotated as important implicitly, namely those that received more attention based on gaze. *Implicit annotation* vectors for each user were constructed as follows. First, as a target number of annotated words for each user, we used the number k of unique words in the same user's explicit annotation. Since the annotation lengths can also vary substantially for different users, to have comparable results we used the same overlap measure (Formula 4.4.1) that we used for the calculation of explicit agreement. We hypothesised that important answer words would receive more user gaze than others. Gaze on a word was measured based on the amount of time that a user looked at it (total fixation duration), and the number of regressions to this word. While individual differences may lead users to view some specific words for longer duration (for instance, rare or unknown words), we conjecture that words that are of interest in relation to the specific question being answered would benefit from more consistent attention across users.

The top k words used for implicit annotation were extracted first sorting by decreasing regression count, and then sorting by decreasing fixation duration to resolve ties. In Table 6, overlap scores of implicit annotations are displayed. The agreement difference for correct and incorrect answers is similar to explicit annotations and the overlap for correct answer is statistically significantly higher.

4.4.3 Comparison of explicit annotation and user gaze. To investigate how words explicitly annotated by users overlap with words

Table 6: Overlap score between implicit annotation.

Answer quality	User settings: all	without highlighting	with highlighting	between 2 settings
ALL	0.59	$0.67^{\dagger}$	$0.56^{\dagger}$	0.57
INCORRECT	0.58^{*}	$0.65^{\dagger *}$	$0.56^{\dagger}$	0.55^{*}
CORRECT	0.61^{*}	$0.7^{\dagger *}$	$0.56^{\dagger}$	0.6^{*}
LQ & HQ	0.62 & 0.59	$0.72^{\dagger}$ & $0.68^{\dagger}$	$0.57^{\dagger}$ & $0.54^{\dagger}$	0.6 & 0.59

** significant difference between correct/incorrect answer groups;*
† between users shown/not shown automatic highlighting ($p < 0.05$).

that are implicitly determined to be important, we first filtered out all stopwords, due to them being unimportant for comparison. We then compared the explicit user annotations with the implicit annotations. Since the lists were of the same length, we use the Jaccard coefficient as a similarity metric. The first part of Table 7 shows the mean Jaccard scores for all documents calculated between explicit and implicit word lists for each user. Overall, although agreement is not very high on average, there is still some intersection of words. As explained previously, there are many factors that can influence a user's gaze. To account for individual noise, such as long fixations on words that happen to be unknown for a particular user, we also compared *average user* word annotation lists.

Average user word annotation lists (either explicit or implicit) are created by taking the words from all individual lists (constructed as described above), and sorting by decreasing frequency (reflecting how many users have a word in their individual sets). After that, we select the top m most popular words, where m is the mean number of words that were chosen for explicit annotating across all users.

The average similarity scores for all answers between the average user word lists are displayed in the lower part of Table 7. The mean lists for each user group were created only from lists of users from that group. It can be seen that when the data from a greater number of users is aggregated, the more similar the final lists are (Jaccard coefficient between explicit and implicit mean lists is 0.361 for all users, while in divided equal users' parts it is only 0.309 and 0.328). The overlap for averaged words is on average higher than at the individual level.

The agreement between explicit and implicit annotating is higher for correct answers. This could signify that *users are likely to understand explicitly which words indicate that an answer is correct. Conversely, their gaze is distributed more randomly when evaluating incorrect answers, suggesting that there aren't specific or consistent phrases that flag the incorrectness of an answer.* This is also supported by a higher number of areas of interest, and lower explicit and implicit overlap scores, between focus words for different users when dealing with incorrect answers.

5 IMPACT OF HIGHLIGHTING WORDS

The highlighting of certain terms to guide users was intended to assist in the evaluation of answers. In this section we first study differences in the evaluation process between two groups: users who were or were not shown highlighted words. Next, we compare the similarity of automated guiding highlighting to explicit and implicit user annotations. Finally, we study differences in the

Table 7: Similarity of explicit and implicit annotating.

Answer quality	User settings: all	without highlights	with highlights
average of Jaccard coefficient for each user			
ALL	0.284	$0.321^{\dagger}$	$0.247^{\dagger}$
INCORRECT	0.271^{*}	$0.306^{\dagger *}$	$0.235^{\dagger *}$
CORRECT	0.299^{*}	$0.339^{\dagger *}$	$0.259^{\dagger *}$
Jaccard coefficient between "average user word list"			
ALL	0.361	0.309	0.328
INCORRECT	0.337	0.304	0.301^{*}
CORRECT	0.389	0.315	0.36^{*}

* *significant difference between correct/incorrect answer groups;*
$\dagger$ *between users shown/not shown automatic highlights* ($p < 0.05$).

evaluation between those answers where similarities between the highlights and annotations were high and low.

5.1 Assessment Quality

By experimental design, users were randomly distributed into two groups that had very little differences in demographics, education, and English level. We also compared the responses to the pre-task questionnaire items, namely the mean values of interest, familiarity, and perceived difficulty regarding questions. The scores were consistent between the two groups. Thus, we conclude that *the main factor that could influence their answer evaluation process was the presence or absence of automatically highlighted words in an answer.*

Table 2 shows that the overall accuracy of evaluation between the two groups (with or without automatic highlighting of words) is comparable. While users who saw highlighting evaluated high-quality correct answers more accurately, they sometimes mis-rated incorrect answers with low-quality correct answers. Regarding fine-grained answer quality aspects, the difference of ratings between the two groups was not statistically significant. *In other words, we can say that automatically highlighting words does not influence the accuracy of assessment of non-factoid answers.*

5.2 Important Word Annotation

The three last columns of Table 5 show the extent to which people shown/not shown automatic highlighting agreed when marking a set of important words in an answer, both within and between the two user groups. There is good agreement between users within their groups; on average, users who were shown highlighting have slightly higher agreement to users who did not see it. The agreement was the lowest between users in two different groups, especially, for incorrect answers. Manually inspected cases showed that users who saw highlighted words in an incorrect answer had a tendency to annotate them as negative. It also explains higher agreement within this group for incorrect answers. At the same time, these assessors sometimes skipped annotating of those already highlighted words in correct answers, which is supported by the fact that the average count of annotated words in this group was lower than in the group without automatic highlighting, 7.8 versus 11.0.

In contrast, the agreement of the implicit vector of important words (the construction of which was described in the previous subsection) for users in the group with automatic highlighting was lower than in the group without highlighting (Table 6). This was a counter-intuitive result as we expected users would pay attention to the words which were specifically highlighted in the text. *It appears that users read highlighted words faster on average, and returned to them less frequently, studying other words instead.*

The agreement between explicit and implicit annotations in the group with automatic highlighting was also lower than for the group without. We can therefore conclude that, *despite the fact that automatic highlighting does not impact the accuracy of the answer evaluation, it can influence what parts of the answer users pay attention to during evaluation.* However, as we stated earlier, we could not find any significant differences in the correctness ratings between the groups.

5.3 Differences In Gaze Metrics

We compare gaze metrics between the two different user groups. As shown in Table 4, users from the groups with automatic word highlighting rated answer quality while spending less time on the answer screen. The difference between the two groups was significant, as was the difference for mean fixation durations and regression durations. While fixation and regression counts were not significantly different between the groups, mean fixation and regression durations and counts were lower for the group with highlighting, which could indicate that these users required less effort for answer evaluation. We also report all metrics for both groups with exclusion of all assessment errors. The trends show that users with automatic highlighted word rated an answer faster and with fewer fixations and regressions even when considering correct labeling only. *This finding supports the hypothesis that highlighting important words in an answer for a non-factoid question makes the evaluation process easier for a user.*

5.4 Similarity of Highlights and Annotations

To investigate the similarity of automatic highlighting (based on *TF-IDF* and capitalized words) and explicit or implicit user annotations, and how this influences evaluation, we first need to construct a *target user annotation* vector. As users who saw suggested highlighting are biased in their implicit and explicit word lists, we exclude such users from comparisons for the current analysis. Moreover, since the highlighting aims to help a user with evaluation, we only use the feedback of those users who correctly rated answers; otherwise we could end up with the misleading word suggestions. As an explicit or implicit user word annotation we used an "average user word list", which was described in Section 4.4.3. In the remainder of this analysis, we refer to these average user word lists, constructed only over data from users who did not see automatic word highlighting, as "*implicit target user annotating*" and "*explicit target user annotating*"

The first column of a top part of Table 8 shows the similarity calculated as the Jaccard coefficient between automatic word highlighting based on *TF-IDF* and explicit user annotations. The similarity is the highest for high quality answers which means it is mostly not capital or *TF-IDF* heavy words that allow people to understand the incorrectness of an answer. However, as we previously showed, the agreement on words among users is higher for correct documents as well. In the same way, we compared the shown highlighting with implicit user annotations. The results are

reported in the first column of Table 8 on the bottom. Generally, the similarity is higher than with explicit annotation and this difference is statistically significant overall and among incorrect answers (Student's t-test, $t - statistic = 3.9$, $p < 0.01$). This could be explained by the fact that users could not always explicitly identify (and annotate) the complete list of important words which they implicitly paid attention to during the rating process. We observed this in the previous subsection as there was a low agreement between explicit and implicit word vectors for incorrect answers.

We further investigate how this similarity is connected with the user's speed of evaluation. As we have shown, users from the group with automatic word highlighting evaluated answers statistically significantly faster than users who did not see highlighting. We found statistically significant weak negative correlation (Pearson -0.27, $p = 0.02$) between the total view time of these users and the similarity of automatic word highlighting to the explicit user annotations. We also found low negative correlation of the view time with similarity to implicit words (Pearson -0.14, $p < 0.001$). For completeness, we checked the correlation between the Jaccard scores and length of suggestions to reject the hypothesis that the quality of automatic word highlighting depends on their length; the correlation was not statistically significant (Pearson -0.05 , $p = 0.653$). We also calculated the average similarity for the subset of answers where total view time results were inconsistent with the general trend (the outliers which were viewed longer by the users who were shown an answer with automatic word highlighting). Notably, their similarity with target explicit annotations for these answers is slightly lower with the value of 0.254, versus 0.275 for those answers that were rated more quickly by users who were shown automatic word highlighting. However, this difference is not statistically significant. We can conclude that *the more similar automatic word highlighting is to users' annotations, the more helpful it is for them when evaluating answers.*

6 ANALYZING TRANSFORMER ATTENTION

The aim of this section is to understand the level of similarity between words obtained from an attention map of a Transformer model [33] fine-tuned on the non-factoid question evaluation task and words explicitly or implicitly annotated by users in our study. Here we firstly explain how we extracted sets of important words from Transformer model attention maps. After that, we give details on the particular model we used. Finally, we compare Transformer model attention with implicit or explicit user annotations and "baseline" highlighting we used in our study.

6.1 Attention Construction

Self-attention assigns a weight (attention) from each word in a sentence to each other word, which can be interpreted as word importance and transformed into scores for a word highlighting algorithm. In our case, we input the query and answer simultaneously into a Transformer neural network, obtaining query-to-answer and answer-to-answer attention maps. Then, we calculate the importance of each answer token with respect to question tokens by averaging attention weights leading to this token from the question. We compute answer-to-answer token importance in the same fashion. Finally, we summed the aforementioned averages to form the attention score for each token:

$$attention\_score(t) = \frac{\sum_{q \in Q} w_{qt}}{|Q|} + \frac{\sum_{a \in A, a \neq t} w_{at}}{|A|}$$

While our user study yielded data of a form appropriate for training a model, the quantity of labels is insufficient; we therefore use a pre-trained one, specifically a "large" uncased BERT model with whole word masking [10], which has 24 layers, each containing 16 attention heads. The BERT model was chosen due to its success in the similar task of a non-factoid answer ranking [22]. We study the average and standalone performance of attention maps of the last layer heads, as they have no shared parameters between them.

Given a BERT model, we constructed the input as **[CLS] <question> [SEP] <answer> [SEP]**, where **[CLS]** and **[SEP]** are special tokens that indicate the beginning of the input and the separation between the sentences, respectively. Using this form of input, we obtained attention maps for each question-answer pair. We excluded weights that were connected to the **[CLS]** and **[SEP]** special tokens, as they do not appear in user explicit or implicit annotation vector, and Clark et al. [6] demonstrated that these tokens have insignificant impact on the accuracy for most heads. We leveraged the BertVis visualization tool Vig [34] to extract attention maps from the model. Additionally, we wanted to infer the attention score for word-level tokens, but the BERT model uses byte-pair encoding as its tokenization method [28], leading to some words being split into sub-words. Thus, we had to deconstruct the BERT text representation by merging tokens starting with the special symbol '##' and summing their attention scores to get an attention score for the original word, as was proposed by Clark et al. [6]. The final vector could be interpreted as an average importance weight for every word in an answer.

To extract more informative attention maps, we fine-tuned the base BERT model to the non-factoid QA evaluation task similar to the one users in our study were asked to do. Given a QA pair from the nfL6 dataset, the model had to predict if an answer was correct or not. First, we excluded QA pairs that were used in our user study since they also came from this dataset. Then, we randomly chose 1% of the rest as holdout test data, and divided the remainder into training and validation data (80% and 20%, respectively). Then we fine-tuned the model for 2 epochs reaching 0.96 accuracy on the test set. After that, we discarded the last classification layer, and used the remaining layers to obtain attention maps for held out subset of QA pairs - the same as we used in our user study.

The word highlighting task can be formulated in a way similar to our user study task as annotating n important words from an answer for a non-factoid question. To construct automatic highlighting from BERT attention weights, we firstly need to predict n. We chose the average count of words that were explicitly annotated by users as the target number of highlighted words. However, such annotations may not be available for new texts. Thus, we predicted word annotation counts using linear regression, with answer length as the only feature (the correlation between the number of highlighted words and answer length is 0.7, Pearson, $p < 0.001$).

For a given answer, we therefore predicted n using regression, and then selected the top-n words in the answer with the highest attention scores. We used an average of attention maps from all heads of the Transformer layer as the overall BERT attention.

Table 8: Similarity between model and user annotations.

Answer quality	Suggestion type: shown TF-IDF	base BERT	fine-tuned BERT
*Jaccard coefficient with user **explicit** annotations*			
ALL	0.266	0.276	0.312
INCORRECT	0.274	0.256	0.269
CORRECT	0.261	0.299	**0.363***
LQ & HQ	0.236 & 0.288	0.289 & 0.313	0.341 & 0.391
*Jaccard coefficient with user **implicit** annotations*			
ALL	0.295	0.314	0.316
INCORRECT	0.299	0.292	0.278
CORRECT	0.29	0.339	**0.361***
LQ & HQ	0.274 & 0.314	0.347 & 0.328	0.372 & 0.346

** significantly different from TF-IDF ($p < 0.05$).*

6.2 Comparison With User Annotations

To evaluate whether BERT and users in our study pay attention to the same words, we calculated the Jaccard similarity with explicit and implicit user annotations by the same way we did such comparison with shown TF-IDF highlighting in Section 5.4. The results are shown in Table 8. Overall, there is much better overlap with target user annotating comparing to the TF-IDF baseline. The difference between *TF-IDF* and fine-tuned BERT is statistically significant for correct answers ($p < 0.004$). Differences for all and incorrect answers are not statistically significant ($p = 0.065$ and $p = 0.83$, respectively). BERT attention similarity is higher compared to *TF-IDF* for correct answers, and lower for incorrect answers. It appears that weights for words signaling that the answer is incorrect could not be obtained from the average attention of all heads. We investigate this in the next subsection.

6.3 One Head Is All You Need

Previous research shows that in BERT models, only a small subset of heads is important for the target task [23, 36]. Inspired by this, we investigated whether we could use the attention map from a single "best" head on the last layer to obtain a final score that is at least not worse than from the average head approach reported earlier. Recall that the performance of our method is different for incorrect and correct documents; we therefore analyzed them separately. Since the number of available documents is quite small (43 incorrect and 37 correct), we used leave-one-out cross-validation for the selection of the best head. For each document, we selected the best performing head on all documents except the current one. Then we calculated the similarity for that document using the selected head. The comparison between best-head performance and averaged attention score is reported in Table 9.

Table 9: Best-head Jaccard coefficient for *fine-tuned* BERT.

Answers	With explicit: best head	average	With implicit: best explicit head	best head	average
INCORRECT	**0.292** 11 head	0.269	**0.284** 11 head	**0.313** 0 head	0.278
CORRECT	**0.371** 13 head	0.363	**0.366** 13 head	**0.381** 5 head	0.361

Notably, for incorrect and correct answers there was only one best head over all folds: head 13 for correct answers, and head 11 for incorrect answers. Six out of sixteen heads always perform worse than the average. The performance with head 11 and head 13, which give the best similarity with explicit highlighting, also give the second-best similarity with implicit annotations. The improved similarity for incorrect documents with head selection also becomes higher than baseline *TF-IDF* highlighting: 0.292 versus 0.274 for explicit incorrect, and 0.313 versus 0.299 for implicit incorrect. The overall improvement is promising and relatively stable across folds on average, indicating that, given explicitly annotated answers, we can select the best head and improve the method scores.

6.4 Implications of Highlighting

From our findings so far, we can conclude that providing automatic word highlighting accelerates the user rating of answer correctness in the challenging task of non-factoid QA evaluation. Specifically, it leads to lower total view time, fixation count, and regression count. However, the automatic word highlighting proposed by Qu et al. [25] only moderately correspond to the words that users annotated as important, sometimes even misleading them in our study. To improve over the baseline, we can consider the use of a more complex algorithm that makes use of self-attention - a key mechanism in state-of-the-art Transformer neural networks that allows them to track dependencies between words. As we have shown, highlighting based on the attention map of a fine-tuned BERT model is more similar to human gaze and explicit annotations and, hence, could be potentially used during the non-factoid QA evaluation process or as highlighting in answer snippets on a SERP.

7 CONCLUSION

In this paper, we studied how users interact with answers of different quality levels that are presented in response to non-factoid questions submitted to web search systems, examining important words that are explicitly annotated by users, and words that are the focus of user attention based on gaze from eye-tracking patterns.

Our first research question focused on the challenging process of evaluation of a passage-level answer to a non-factoid question. The results show that user's interest in the question and perceived question difficulty impact rating accuracy. Also, answers were considered to be correct only when they were also complete. Thus, an answer to a non-factoid question should not only contain correct information but also give enough of it to cover most of the question's aspects, such as proper explanations, several opinions, examples, and so on. Most of the answers selected as "best" from the (Yahoo!Answers) CQA platform (that we used in our study) did not satisfy users with low levels of knowledge about the question topic. It should be taken into account when using CQA platform data for non-factoid QA. According to user annotations, it is harder for people to identify incorrect parts of a passage-level answer than correct ones, which is supported by higher agreement between explicit and implicit annotations of a user for correct answers. The same holds for the agreement of annotations between users. This indicates that people understand well why an answer is correct. Conversely, for incorrect answers user gaze is distributed more randomly, with a higher number of areas of interest and lower explicit

and implicit annotation overlap between users. Nevertheless, we found that users spend more time and effort to understand that an answer is correct than that it lacks some information or is incorrect.

The next research question addressed automatic highlighting of important words in answers. The results demonstrated that it helps users to rate answers more quickly and with less effort while maintaining the same level of quality. However, highlighting changes implicit and explicit annotations from users, drawing their attention more to the words that are not highlighted. The overall similarity of the *TF-IDF* highlighting we utilized in the study and explicit or implicit user annotations is not very high. Despite that, cases with higher similarity were evaluated faster than those with lower similarity, suggesting that highlighting is useful when it models user annotations well.

Our third research question was to study whether the attention mechanism of a Transformer neural network assigns weights to words similarly to explicit or implicit user annotations on important words. The results of our experiments showed that the term highlighting, constructed from the weights of a fine-tuned BERT model, is more similar to users' annotations than the baseline method that selects terms based on *TF-IDF* weights and capitalisation.

Overall, our findings and the suggested highlighting method can be used to make the process of non-factoid QA evaluation easier. This approach could also be used to investigate how modifying the highlighting of snippets for answers for non-factoid questions influences user interaction with search result pages overall, and potentially improve that experience. While our experiments demonstrated how term highlighting that is highly aligned with user explicit or implicit term annotation improves user evaluation, our current experimental data cannot demonstrate whether *TF-IDF* or BERT highlighting is more effective as there is needed a separate user study, giving an avenue for future work.

REFERENCES

[1] Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34, 4 (2008), 555–596.

[2] Jane Ashby, Keith Rayner, and Charles Clifton. 2005. Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology* 58, 6 (2005), 1065–1086.

[3] Lorena Leal Bando, Falk Scholer, and Andrew Turpin. 2010. Constructing query-biased summaries: a comparison of human and system generated snippets. In *In Proc. of IIiX*. ACM, 195–204.

[4] Andrei Broder. 2002. A taxonomy of web search. *ACM Sigir forum* 36, 2 (2002), 3–10.

[5] Lydia B. Chilton and Jaime Teevan. 2011. Addressing People's Information Needs Directly in a Web Search Result Page. In *In Proc. of WWW*. 27–36.

[6] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proc. ACL*. 276–286.

[7] Charles LA Clarke, Eugene Agichtein, Susan Dumais, and Ryen W White. 2007. The influence of caption features on clickthrough patterns in web search. In *In Proc. of SIGIR*. 135–142.

[8] Daniel Cohen and W Bruce Croft. 2016. End to end long short term memory networks for non-factoid question answering. In *In Proc. of ICTIR*. ACM, 143–146.

[9] Edward Cutrell and Zhiwei Guan. 2007. What are you looking for?: an eye-tracking study of information usage in web search. In *Proc. SIGCHI*. 407–416.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *In Proc. of NAACL*. 4171–4186.

[11] Laura A Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *In Proc. of SIGIR*. ACM, 478–479.

[12] Jacek Gwizdka. 2014. Characterizing relevance with eye-tracking measures. *In Proc. of IIiX* (08 2014). https://doi.org/10.1145/2637002.2637011

[13] Katja Hofmann, Lihong Li, Filip Radlinski, et al. 2016. Online evaluation for information retrieval. *FnTIR* 10, 1 (2016), 1–117.

[14] Tereza Iofciu, Nick Craswell, and Milad Shokouhi. 2009. Evaluating the impact of snippet highlighting in search. *UIIR-2009* (2009), 44.

[15] Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. 2014. A neural network for factoid question answering over paragraphs. In *In Proc. of EMNLP*. 633–644.

[16] Thorsten Joachims and Filip Radlinski. 2007. Search engines that learn from implicit feedback. *Computer* 40, 8 (2007), 34–40.

[17] Johanna K Kaakinen and Jukka Hyönä. 2007. Perspective effects in repeated reading: An eye movement study. *Memory & Cognition* 35, 6 (2007), 1323–1336.

[18] Mostafa Keikha, Jae Hyun Park, W Bruce Croft, and Mark Sanderson. 2014. Retrieving passages and finding answers. In *In Proc. of ADCS*. 81.

[19] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. In *In Proc. of EMNLP-IJCNLP*. 4365–4374.

[20] Dmitry Lagun, Chih-Hung Hsieh, Dale Webster, and Vidhya Navalpakkam. 2014. Towards Better Measurement of Attention and Satisfaction in Mobile Search. In *In Proc. of SIGIR*. 113–122.

[21] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology* 59, 7 (2008), 1041–1052.

[22] Yosi Mass, Haggai Roitman, Shai Erera, Or Rivlin, Bar Weiner, and David Konopnicki. 2019. A Study of BERT for Non-Factoid Question-Answering under Passage Length Constraints. *ArXiv* abs/1908.06780 (2019).

[23] Paul Michel, Omer Levy, and Graham Neubig. 2019. Are Sixteen Heads Really Better than One?. In *Proc. NeurIPS*. 14014–14024.

[24] Jae Hyun Park and W Bruce Croft. 2015. Using key concepts in a translation model for retrieval. In *In Proc. of SIGIR*. 927–930.

[25] Chen Qu, Liu Yang, W. Bruce Croft, Falk Scholer, and Yongfeng Zhang. 2019. Answer Interaction in Non-factoid Question Answering Systems. In *Proc. CHIIR*. 249–253.

[26] Keith Rayner, Erik D Reichle, Michael J Stroud, Carrick C Williams, and Alexander Pollatsek. 2006. The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and aging* 21, 3 (2006), 448.

[27] Denis Savenkov, Pavel Braslavski, and Mikhail Lebedev. 2011. Search snippet evaluation at yandex: lessons learned and future directions. In *Proc. CLEF*. 14–25.

[28] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *In Proc. of ACL*. 1715–1725.

[29] K Spärck Jones, Steve Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management* 36, 6 (2000), 809–840.

[30] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online QA collections. In *In Proc. of ACL-08: HLT*. 719–727.

[31] Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018. An Analysis of Attention Mechanisms: The Case of Word Sense Disambiguation in Neural Machine Translation. In *In Proc. of WMT 2018*. 26–35.

[32] Anastasios Tombros and Mark Sanderson. 1998. Advantages of query biased summaries in information retrieval. In *In Proc. of SIGIR*. ACM, 2–10.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proc. NIPS*.

[34] Jesse Vig. 2019. A Multiscale Visualization of Attention in the Transformer Model. In *In Proc. of ACL*. 37–42.

[35] MK Vijaymeena and K Kavitha. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal* 3, 2 (2016), 19–28.

[36] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. In *In Proc. of ACL*. 5797–5808.

[37] Jaap Walhout, Paola Oomen, Halszka Jarodzka, and Saskia Brand-Gruwel. 2017. Effects of task complexity on online search behavior of adolescents. *Journal of the Association for Information Science and Technology* 68, 6 (2017), 1449–1461.

[38] Kyle Williams, Julia Kiseleva, Aidan C. Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. 2016. Is This Your Final Answer? Evaluating the Effect of Answers on Good Abandonment in Mobile Search. In *Proc. SIGIR*. 889–892.

[39] Gary S Wolverton and David Zola. 1983. The temporal characteristics of visual information extraction during reading. In *Eye movements in reading*. 41–51.

[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 *In Proc. of ECIR*. 115–128.

[41] Liu Yang, Minghui Qiu, Swapna Gottipati, Feida Zhu, Jing Jiang, Huiping Sun, and Zhong Chen. 2013. Cqarank: jointly model topics and expertise in community question answering. In *In Proc. of CIKM'13*. 99–108.

[42] Yukun Zheng, Jiaxin Mao, Yiqun Liu, Zixin Ye, Min Zhang, and Shaoping Ma. 2019. Human Behavior Inspired Machine Reading Comprehension. In *Proc. SIGIR*. 425–434.