Answer Interaction in Non-factoid Question Answering Systems

Chen Qu, Liu Yang, W. Bruce Croft
University of Massachusetts Amherst
{chenqu,lyang,croft}@cs.umass.edu

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

Yongfeng Zhang
Rutgers University
yongfeng.zhang@rutgers.edu

ABSTRACT

Information retrieval systems are evolving from document retrieval to answer retrieval. Web search logs provide large amounts of data about how people interact with ranked lists of documents, but very little is known about interaction with answer texts. In this paper, we use Amazon Mechanical Turk to investigate three answer presentation and interaction approaches in a non-factoid question answering setting. We find that people perceive and react to good and bad answers very differently, and can identify good answers relatively quickly. Our results provide the basis for further investigation of effective answer interaction and feedback methods.

KEYWORDS

User Interaction; Answer Interaction; Answer Presentation; Non-factoid Question Answering; Information-seeking

ACM Reference Format:
Chen Qu, Liu Yang, W. Bruce Croft, Falk Scholer, and Yongfeng Zhang. 2019. Answer Interaction in Non-factoid Question Answering Systems. In *2019 Conference on Human Information Interaction and Retrieval (CHIIR '19), March 10–14, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3295750.3298946

1 INTRODUCTION

Classic information retrieval (IR) systems aim to return a list of relevant documents on a search engine result page (SERP). This type of presentation is often described as "ten blue links", because users typically need to click on the ranked results and be redirected to the documents. Modern search engines have paid attention to search results diversification [3, 21] and heterogeneous content presentation [29]. Recently, several works have focused on retrieving extractive answers instead of documents [11, 18, 35, 37, 38]. Industrial examples include Google's *featured snippets*,[1] which display a potential answer extracted from the top search result.

If search engines can return a list of potential answers rather than documents, it is essential to study the most effective way to present these answers and interact with users. Specifically, this research question should be emphasized in non-factoid question answering (QA) systems. This is because non-factoid QA poses unique challenges to answer presentation and interaction as it requires several answer sentences or passages, instead of simple entity-based answers as in factoid QA.

User interaction with SERPs has been widely studied using search logs that contain clicks and query reformulations [2, 23]. Furthermore, other works focus on interaction and feedback methods for document retrieval by studying real users instead of search logs [12, 13]. However, fine-grained presentation and interaction processes with answers have rarely been investigated in previous work. Additional information is needed from observing what constitutes a good answer when users provide fine-grained and precise feedback, instead of simply indicating whether the answer is relevant or not. We believe that studying fine-grained user interaction and feedback can lead to more effective answer finding, as well as having an impact on the design of conversational search systems.

In this work, we investigate three answer presentation and interaction approaches (Line by Line, Passage Highlight, and Passage Highlighting with Suggested Words) to understand how people perceive good and bad answers. The Line by Line setting reveals a potential answer passage one line at a time and observes people's reactions as they go through the passage. The Passage Highlight setting presents the full passage, and instructs users to highlight important words that make them believe the passage is a good or bad answer. The third setting is built upon the second one, and includes some suggested words emphasized with special styles. We hired crowdsourcing workers from Amazon Mechanical Turk (MTurk)[2] to conduct the experiments. Based on these fine-grained experiments, we find that people perceive good answers and bad answers very differently, which could lead to more effective relevance feedback schemes. For example, people do not hesitate to rate a bad answer, but they can be severe on the answer quality judgments even in some cases where the passage is the answer. Another finding is that people's initial impressions of answer quality are usually correct, and they become more and more confident about answer quality as they go through the answer. In addition, we investigate the relation between answer quality and QA text similarity and find that they are not always correlated.

Our contributions can be summarized as follows. (1) We conduct one of the first fine-grained analyses on answer presentation and interaction in a non-factoid QA setting. (2) We provide an empirical analysis to answer an important research question: is answer quality related to QA text similarity? Our findings can be used to design a more interactive IR system that emphasizes answer retrieval. In addition, our work also has implications for conversational search, since it is essentially a multi-turn interaction process.

2 RELATED WORK

User Interaction and Relevance Feedback. Relevance feedback [4, 8, 12, 15, 39] is an important and early interactive method in IR systems. In practice, pseudo-relevance feedback (PRF) [6, 14, 16]

[1] https://support.google.com/webmasters/answer/6229325

[2] https://www.mturk.com/

CHIIR '19, March 10–14, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-6025-8/19/03…$15.00
https://doi.org/10.1145/3295750.3298946

is widely used. It assumes that top-ranked documents are relevant and thus can be used for query expansion. In contrast to PRF, some approaches focus on explicit interactions with the user [1, 7, 13, 17]. This work builds on these early papers and focuses on explicit and fine-grained user interaction with answers. Our methods could be especially useful when the interaction bandwidth is limited, such as in mobile search and conversational search.

Answer Retrieval. IR systems are evolving from document retrieval to answer retrieval. Substantial work has been done in factoid QA [9, 22, 34, 37, 38], community QA [10, 24, 24, 25, 32, 33], and non-factoid QA [5, 11, 18, 35]. All these works focus on finding effective methods for answer retrieval. However, even the most effective method can occasionally fail to find the answers. In that case, it is essential to employ user interaction and feedback methods to retrieve the answer in an iterative manner. In this work, we study answer presentation and interaction techniques in a non-factoid QA setting, as an essential complement to answer retrieval models.

Information-seeking Conversations. An information-seeking conversation typically involves multiple turns of interaction and information exchange between an information seeker and provider. Radlinski and Craswell [20] described a theoretical framework for conversational search and desirable properties in such systems. In addition, several works [19, 26, 27, 30, 31] observed and studied information-seeking conversations between humans and addressed different facets of such interactions. Finally, conversational recommendation [40] and response ranking [36] have been explored under this multi-turn interaction setting.

3 OUR APPROACH

3.1 Overview

We conduct an observational study of how people perceive answer quality under different interactive settings. This can help us identify effective methods for answer presentation and interaction. In this task, people are given a question and a short passage. The passage may or may not be a good answer to the question.[3] We present the answer passage in three ways, namely, Line by Line, Passage Highlight, and Passage Highlighting with Suggested Words. These settings are designed to obtain fine-grained user feedback to test various answer presentation and interaction methods.

3.2 Line by Line

In this setting, the answer passage is presented line by line. One line is typically one sentence. At each line, we instruct the annotators to give a rating of how confident they are that the passage is or contains a good answer to the question. This rating is based on the current line and previous lines. Lines that follow the current line are hidden. The confidence rating is provided on a scale of -2 to 2:

- -2: Confident this passage is not an answer to the question.
- -1: Believe that this passage might not be an answer.
- 0: Not sure yet.
- 1: Believe that the passage might be an answer to the question.
- 2: Confident this passage is an answer to the question.

This setting is designed to observe the evaluation of answer quality as people go through a potential answer.

[3] "Good" refers to good quality. Verifying the facts in the passages is not required.

3.3 Passage Highlight

In this setting, we present the full answer passages and instruct the annotators to highlight positive and negative words or phrases in the passage (sentences are not encouraged). The *positive words* are those that help to convince the annotators that the passage is a good answer. For example, these words may present a specific answer or introduce key arguments. In contrast, the *negative words* make the annotators feel the answer is of bad quality. For example, these words can be indicators of irrelevant issues, or may reveal that the answer providers are uncertain about their answers. Annotators are instructed to highlight complete words only. At least one highlight for each passage needs to be made for a successful submission. Figure 5 gives an illustration for the highlighting interface.[4] In addition, annotators are asked to give a rating on overall answer quality. The answer quality can be chosen on a scale of 0 to 2:

- 0: It is not an answer to the question.
- 1: It is an answer to the question, but not of good quality.
- 2: It is a good answer to the question.

This setting is designed to obtain fine-grained feedback on answer quality evaluation and observe rating agreement.

The main challenge of this setting is to quantify annotators' agreement on highlights. Although the annotators are instructed to highlight whole words only, they sometimes highlight partial words. So first, the character-level highlights need to be transformed into word-level highlights. Then a *consistency rule* is applied to highlight all occurrences of a word if this word is highlighted by the annotator. This rule is not applicable to stop words. Then the next step is transforming the highlighted passage into a string by denoting the highlighted words as "1" with others as "0". In this way the agreement of highlights can be cast into a string similarity problem. The overlap coefficient [28] is used to tackle this problem:

$$\text{overlap}(H_1, H_2) = \frac{|H_1 \cap H_2|}{\min(|H_1|, |H_2|)} \tag{1}$$

where H_1 and H_2 are string representations of highlights from two annotators. This method can compute agreement among multiple annotators. However, since complete agreement with more than two annotators is rare due to the open style of the task, only pairwise agreement is computed. The largest value among all pairwise agreements is considered as the agreement for this QA pair. We adopt this setting because perceptions and highlights of answer key words are highly subjective.

3.4 Passage Highlighting with Suggested Words

This setting is very similar to the previous setting. The only difference is that we mark some suggested words in the passage for the annotators' reference. These suggested words are marked in bold font and blue color to draw user attention. Figure 5 gives an example of the highlighting interface. The suggested words meet one of the following criteria: (1) Words that start with a capital letter, such as acronyms or proper nouns. Words that start a sentence have been excluded. (2) Words that have the top five tf-idf value in this passage. Idf values are computed with a Wikipedia dump (date: 20180520). The annotators understand that they do not have to keep to the suggested words. This setting is designed to compare reactions with and without the suggested words.

[4] The words in blue are marked for the next setting. They are in black in this setting.

4 EXPERIMENTS

4.1 Data Preparation

We sampled 200 QA pairs from nfL6,[5] a non-factoid community QA dataset. A question typically comes with multiple answers provided by the community, with one of them selected as the accepted answer. Our sampled data contains 100 questions, and each question has a good answer and a bad answer. The good answers are the accepted answers of the questions. To match a bad answer to each question, we first use BM25 to collect a small pool of candidate answers for each question and then use a BLSTM model [5] to rerank. We set a criterion that all the answers can be naturally split into four sentences so that the answers would have appropriate lengths, and to work with a fixed number of lines in the Line by Line setting.

4.2 MTurk Setup

We employed crowdsourcing workers (turkers) through Amazon Mechanical Turk (MTurk) to annotate the QA pairs. The three settings are conducted separately. Each QA pair receives annotations from three different turkers. Turkers conduct annotation in the form of assignments, which contain instructions, annotation examples, quiz questions on the instructions, and five QA pairs to annotate (with a combination of good and bad answers). We only use annotations from turkers who have passed the quiz test in the analysis. In addition, the turkers are required to have a HIT (Human Intelligence Task) approval rate of 95% or higher, a minimum of 1,000 approved HITs, and be located in US, Canada, Australia or Great Britain. The turkers are paid $0.5/assignment.

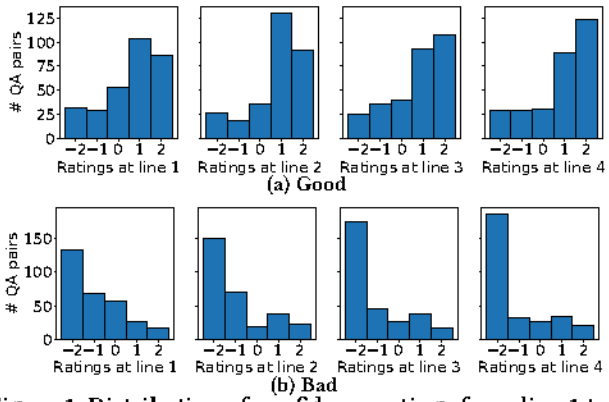
4.3 Line by Line

Figure 1a presents the distribution of confidence ratings from line 1 to line 4 for good answers. The most common confidence rating for line 1 and line 2 is 1. The most common rating gradually shifts to 2 at line 3 and line 4. The distribution of ratings consistently moves up to 2 from line 1 to line 4. Figure 1b presents the distribution of confidence ratings from line 1 to line 4 for bad answers. The most common confidence rating is always -2 through line 1 to line 4. The number of -2 ratings gradually increases as the passage is revealed to the turkers. These indicate that, *for good answers, the turkers have a sense that the answers might be good at the beginning, but they hesitate to make confident ratings until the latter half of the passage is revealed. In contrast, for bad answers, most of the turkers are able to determine the answer quality from the beginning.*

We also use a χ^2 test to evaluate the difference of confidence ratings between the previous line (expectation) and the current line (observation). We observe the same patterns for good answers and bad answers: the shifts of confidence ratings from line $1 \rightarrow 2$ and line $2 \rightarrow 3$ are statistically significant with p-value <0.01, while line $3 \rightarrow 4$ shows an insignificant difference. This indicates that *some people can determine answer quality quickly while others are slower, but they can make a decision before the last line is revealed.*

Figure 2 presents the majority of confidence ratings for all questions. The rating leaps from 1 to 2 between line 2 and line 3 for good answers. The rating remains at -2 throughout the passage for bad answers. This result double confirms the conclusions above.

[5] https://ciir.cs.umass.edu/downloads/nfL6/



(a) Good

(b) Bad

Figure 1: Distribution of confidence ratings from line 1 to 4

In addition to the overall analysis presented above, we also focus on the level of individual QA pairs. Since each QA pair is annotated by three different turkers, we take the majority vote of the ratings at each line as the final rating. We plot the rating trends of individual QA pairs for a better illustration in Figure 3. We only plot the most commonly observed trends with more than two occurrences (in blue or red) to reduce noise. For example, if four QA pairs have the same trend of "$0 \rightarrow 1 \rightarrow 2 \rightarrow 2$", we consider the occurrence of this trend as four. These common trends constitute about half of total trends. We also plot the remaining trends in gray to show that infrequent trends can be very diverse. The line widths are set to the square root of the trend occurrences to demonstrate trend frequency, while avoiding lines being too wide.

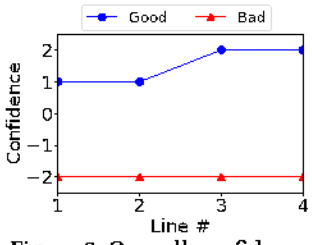


Figure 2: Overall confidence

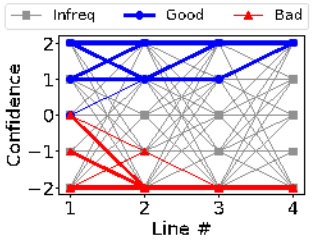


Figure 3: Common trends

As presented in the figure above, the common answer quality ratings start at (0, 1, 2) and converge to 2 for good answers. For bad answers, common ratings start at (0, -1, -2) and converge to -2. In addition, a large portion of good answers have consistent ratings of 2 at all lines. Similarly, a very common trend for bad answers is -2 at all lines. These observations indicate that *initial impressions of answer quality are usually correct, because good answers have positive ratings and bad answers have negative ratings from the very beginning. In addition, people become more and more confident about answer quality as they go through the answer.*

4.4 Passage Highlight

Turkers are instructed to rate the answer quality after highlighting the passage. Figure 4 presents the histogram of rated answer quality. We observe that turkers typically rate 1 or 2 for good answers and 0 for bad answers. This indicates that *turkers do not hesitate to rate*

a bad answer, but they can be severe on the answer quality judgments even in some cases where the passage is the answer.

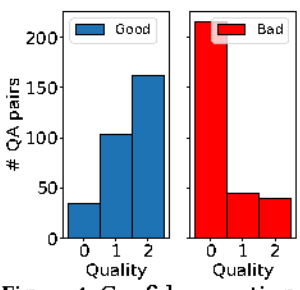


Figure 4: Confidence ratings for the second setting

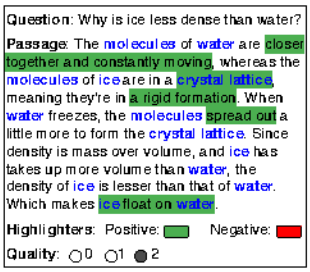


Figure 5: An illustration of the highlighting interface

On average, turkers highlight 7.57 words in good answers and 6.35 words in bad answers. In addition, the consistency rule adds about 0.5 words in both cases. Figure 6 presents the agreement of highlights in four different cases: positive highlight on good answers, positive highlight on bad answers, negative highlight on good answers, and negative highlight on bad answers.

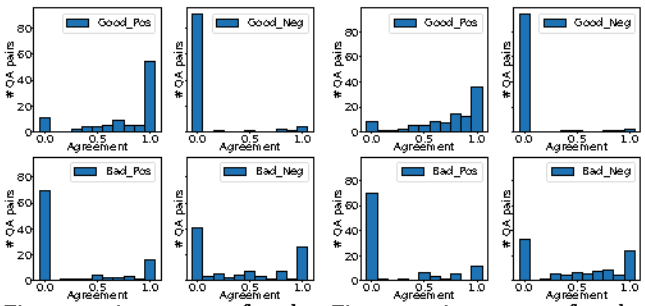


Figure 6: Agreement for the second setting

Figure 7: Agreement for the third setting

We notice the low agreement in Figures "Good_Neg" and "Bad_Pos". This can be accounted for by the fact that turkers rarely use negative highlights for good answers or positive highlights for bad answers. When one or both turkers in an agreement pair do not use a certain type of highlight (positive or negative), the agreement is 0 (turkers need to make at least one highlight, but they do not have to use both types of highlights). Therefore, we focus on positive highlights on good answers and negative highlights on bad answers. In the former case, more than half of the QA pairs achieve complete agreement, which means one turker's annotation is either exactly the same with or a subset of the other's. In the latter case, the agreement is fuzzy but it shows a total agreement in about 30% of the QA pairs. The results indicate that *people tend to get good agreement on what makes a good answer good. In contrast, when deciding what makes a bad answer bad, people tend to have more diverse opinions while still managing to achieve some agreement.*

4.5 Passage Highlighting with Suggested Words

In this setting, the distributions of answer quality ratings are very similar to the last setting, where the suggested words are unmarked. We mark an average of 9.48 suggested words in special styles. The average number of highlighted words by turkers is 6.98 for good answers and 6.06 for bad answers, slightly smaller than the last setting. Overall, the agreement for the last setting (Figure 6) is more polarized while the agreement for this setting (Figure 7) has some distribution weight in the middle. These results indicate that some turkers may consider it unnecessary to highlight the marked words even when they believe the words are positive or negative.

One of the goals of this setting is to observe turkers' behavior under the impact of suggested words. This could be done by quantifying the agreement between highlights and suggested words with overlap coefficient. Figure 8 and 9 shows the results for Passage Highlight and Passage Highlighting with Suggested Words settings. They indicate that *turkers tend to base their decision more on the suggested words when these words are marked in special styles. This suggests that marking important words in answers could influence peoples' decision making process in answer quality evaluation.*

4.6 Answer Quality vs. QA Text Similarity

Some QA systems use the text similarity of questions and answers to perform answer retrieval [5, 35]. However, we show that answer quality is not the same as QA text similarity. We take the majority vote of overall ratings for each QA pair as the final answer quality rating. The rating data comes from the third setting (data from the second setting also gives a very similar result). To compute the text similarity between the question and answer in a QA pair, we obtain the tf-idf representations and the aggregated word embedding representations (the sum of the word embeddings in a passage). Then we calculate the cosine similarity for both representations respectively and compute their harmonic mean as the QA similarity measure. We plot the histogram of QA text similarity under each quality level in Figure 10.

We observe from the figure that the QA text similarity is relatively low in general for non-factoid QA. We further make three observations: (1) Quality level 2 has some high similarity values (such as 0.7 and 0.8), which are rare in quality level 1 and 0. (2) Quality level 0 has more low similarity (such as 0) QA pairs than the other two. (3) All three quality levels have large numbers of medium level similarities. These results indicate that *QA text similarity does not necessarily capture answer quality. They can be positively correlated when it comes to very similar or very different QA pairs. However, text similarity cannot determine answer quality effectively if the QA pair has medium level similarity.* For example, the QA pair shown in Figure 5 only has a text similarity of 0.56, but received quality level 2 ratings from all three turkers. We plan to investigate this in more detail in future work.

5 CONCLUSIONS

In this paper, we studied three different fine-grained answer presentation methods in a non-factoid QA setting. We also discovered that QA text similarity does not necessarily capture answer quality in this setting. Our findings are based on crowdsourcing and thus need to be generalized with caution. Our findings can be used in designing a more interactive IR system that emphasizes answer retrieval. Future work will include verifying our findings and exploring other answer interaction methods in a conversational setting.

ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, NSF IIS-1715095 and ARC DP180102687. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] IJsbrand Jan Aalbersberg. Incremental Relevance Feedback. In *SIGIR'92*, 1992.
[2] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. Learning User Interaction Models for Predicting Web Search Result Preferences. In *SIGIR'06*, 2006.
[3] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying Search Results. In *WSDM'09*, 2009.
[4] Elinor Brondwine, Anna Shtok, and Oren Kurland. Utilizing Focused Relevance Feedback. In *SIGIR'16*, 2016.
[5] Daniel Cohen and W. Bruce Croft. End to End Long Short Term Memory Networks for Non-Factoid Question Answering. In *ICTIR'16*, 2016.
[6] Kevyn Collins-Thompson and Jamie Callan. Estimation and Use of Uncertainty in Pseudo-relevance Feedback. In *SIGIR'07*, 2007.
[7] W. Bruce Croft and Roger. H. Thompson. I^3R: A new approach to the design of document retrieval systems. *JASIS*, 1987.
[8] Fernando D. Diaz. Improving Relevance Feedback in Language Modeling with Score Regularization. In *SIGIR'08*, 2008.
[9] Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal DaumÂĬ III. A Neural Network for Factoid Question Answering over Paragraphs. In *EMNLP'14*, 2014.
[10] Peter Jansen, Mihai Surdeanu, and Peter Clark. Discourse Complements Lexical Semantics for Non-factoid Answer Reranking. In *ACL*, 2014.
[11] Mostafa Keikha, Jae Hyun Park, W. Bruce Croft, and Mark Sanderson. Retrieving Passages and Finding Answers. In *ADCS'14*, 2014.
[12] Diane Kelly and Nicholas J. Belkin. Reading Time, Scrolling and Interaction: Exploring Implicit Sources of User Preferences for Relevance Feedback. In *SIGIR'01*, 2001.
[13] Jürgen Koenemann and Nicholas J. Belkin. A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. In *CHI'96*, 1996.
[14] Victor Lavrenko and W. Bruce Croft. Relevance Based Language Models. In *SIGIR'01*, 2001.
[15] Yuanhua Lv and ChengXiang Zhai. Adaptive Relevance Feedback in Information Retrieval. In *CIKM'09*, 2009.
[16] Yuanhua Lv and ChengXiang Zhai. Positional Relevance Model for Pseudo-relevance Feedback. In *SIGIR'10*, 2010.
[17] Robert N Oddy. *Information Retrieval through Man-Machine Dialogue*. MCB UP Ltd, 1977.
[18] Jae Hyun Park and W. Bruce Croft. Using Key Concepts in a Translation Model for Retrieval. In *SIGIR'15*, 2015.
[19] Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. Analyzing and Characterizing User Intent in Information-seeking Conversations. In *SIGIR'18*, 2018.
[20] Filip Radlinski and Nick Craswell. A Theoretical Framework for Conversational Search. In *CHIIR'17*, 2017.
[21] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. Search Result Diversification. *Found. Trends Inf. Retr.*, 2015.
[22] Aliaksei Severyn and Alessandro Moschitti. Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks. In *SIGIR'15*, 2015.
[23] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum*, 1999.
[24] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to Rank Answers on Large Online QA Collections. In *ACL:HLT'08*, 2008.
[25] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to Rank Answers to Non-factoid Questions from Web Collections. *Comput. Linguist.*, 2011.
[26] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. MISC: A Data Set of Information-seeking Conversations. In *CAIR'17*, 2017.
[27] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. Informing the Design of Spoken Conversational Search: Perspective Paper. In *CHIIR'18*, 2018.
[28] M.K. Vijaymeena and K. Kavitha. A Survey on Similarity Measures in Text Mining. *Machine Learning and Applications: An International Journal*, 2016.
[29] Yue Wang, Dawei Yin, Luo Jie, Pengyuan Wang, Makoto Yamada, Yi Chang, and Qiaozhu Mei. Beyond Ranking: Optimizing Whole-Page Presentation. In *WSDM'16*, 2016.
[30] Mei-Mei Wu and Ying-Hsang Liu. Intermediary's information seeking, inquiring minds, and elicitation styles. *Journal of the American Society for Information Science and Technology*, 2003.
[31] Mei-Mei Wu and Ying-Hsang Liu. On intermediaries' inquiring minds, elicitation styles, and user satisfaction. *Journal of the American Society for Information Science and Technology*, 2011.
[32] Xiaobing Xue, Jiwoon Jeon, and W. Bruce Croft. Retrieval Models for Question and Answer Archives. In *SIGIR'08*, 2008.
[33] Liu Yang, Minghui Qiu, Swapna Gottipati, Feida Zhu, Jing Jiang, Huiping Sun, and Zhong Chen. CQArank: Jointly Model Topics and Expertise in Community Question Answering. In *CIKM'13*, 2013.
[34] Liu Yang, Qingyao Ai, Jiafeng Guo, and W. Bruce Croft. aNMM: Ranking Short Answer Texts with Attention-Based Neural Matching Model. In *CIKM'16*, 2016.
[35] Liu Yang, Qingyao Ai, Damiano Spina, Ruey-Cheng Chen, Liang Pang, W. Bruce Croft, Jiafeng Guo, and Falk Scholer. Beyond Factoid QA: Effective Methods for Non-factoid Answer Sentence Retrieval. In *ECIR'16*, 2016.
[36] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, Jun Huang, and Haiqing Chen. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *SIGIR '18*, 2018.
[37] Scott Wen-tau Yih, Ming-Wei Chang, Chris Meek, and Andrzej Pastusiak. Question Answering Using Enhanced Lexical Semantic Models. In *ACL'13*, 2013.
[38] Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. Deep Learning for Answer Sentence Selection. In *NIPS'14*, 2014.
[39] ChengXiang Zhai and John Lafferty. Model-based Feedback in the Language Modeling Approach to Information Retrieval. In *CIKM'01*, 2001.
[40] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. Towards conversational search and recommendation: System ask, user respond. In *CIKM '18*, 2018.

A ADDITIONAL FIGURES

We include some additional figures to illustrate our findings from above. Figures 8 and 9 shows the agreement between turkers' annotations and the suggested words. The two figures correspond to the Passage Highlight and Passage Highlighting with Suggested words settings respectively. In addition, we plot the histogram of QA text similarity under each quality level in Figure 10.

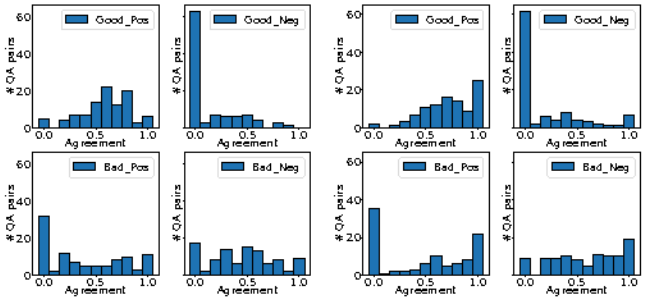


Figure 8: Suggestion overlap for the second setting

Figure 9: Suggestion overlap for the third setting

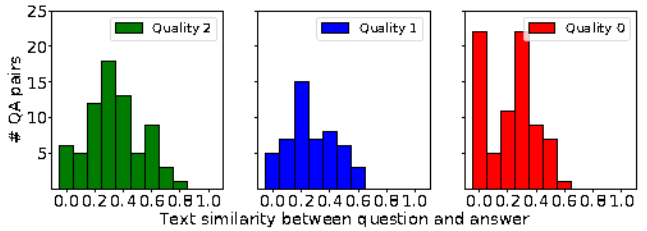


Figure 10: QA text similarity under each quality level