Quantifying Human-Perceived Answer Utility in Non-factoid Question Answering

B. Barla Cambazoglu
RMIT University
Melbourne, VIC, Australia
barla.cambazoglu@rmit.edu.au

Valeriia Baranova
RMIT University
Melbourne, VIC, Australia
lurunchik@gmail.com

Falk Scholer
RMIT University
Melbourne, VIC, Australia
falk.scholer@rmit.edu.au

Mark Sanderson
RMIT University
Melbourne, VIC, Australia
mark.sanderson@rmit.edu.au

Leila Tavakoli
RMIT University
Melbourne, VIC, Australia
leila.tavakoli@rmit.edu.au

Bruce Croft
University of Massachusetts
Amherst, MA, USA
croft@cs.umass.edu

ABSTRACT

Taking a user-centric approach, we study the features that render an answer to a non-factoid question useful in the eyes of the person who asked that question. An editorial study, where participants assess the usefulness of the answers they received in response to their questions, as well as 12 different aspects associated with the answers, indicates considerable correlation between certain aspects such as relevance, correctness, and completeness with the user-perceived usefulness of answers. Moreover, we investigate the effectiveness of some commonly used answer quality measures, such as ROGUE, BLEU, METEOR, and BERTScore, demonstrating that these measures are limited in their ability to capture the aspects of usefulness and have room for improvement. The question answering dataset created in our work is publicly available.

KEYWORDS

non-factoid question answering, perceived answer usefulness, aspect taxonomy, editorial study, answer quality measures

ACM Reference Format:
B. Barla Cambazoglu, Valeriia Baranova, Falk Scholer, Mark Sanderson, Leila Tavakoli, and Bruce Croft. 2021. Quantifying Human-Perceived Answer Utility in Non-factoid Question Answering. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '21), March 14–19, 2021, Canberra, ACT, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3406522.3446028

1 INTRODUCTION

In information retrieval, the quality of a retrieved result document for a given query is usually modeled as a combination of certain aspects of the query and the document, such as the relevance of the document to the query, the popularity of the document, and the importance or authority of the document's source. However, in the context of question answering, where a more direct answer is retrieved for a given question, the aspects that determine the perceived usefulness of an answer may be different. In particular, non-factoid questions present a greater challenge since the usefulness of an answer may depend on the intent of the question.

In this work, we aim to identify the constituents of a high-utility answer in non-factoid question answering, i.e., the aspects that render an answer given to a non-factoid question useful for the question asker. To this end, we first create an initial taxonomy of aspects that may correlate with answer utility, based on an inspection of a sample set of questions and corresponding answers generated in a small-scale editorial study. The initial taxonomy was refined further in a number of steps, resulting in a final set of 12 aspects, which we use to capture the usefulness of a given answer.

Having created our aspect taxonomy, we conduct a multi-step editorial study, involving question/answer generation and usefulness/aspect assessment steps. In the question generation step, all participants act as askers and generate questions of different categories. In the answer generation step, we obtain answers to the generated questions from two different answer sources: participants, who act as answerers, and a commercial web search engine. Finally, in the usefulness/aspect assessment step, the askers assess the usefulness of the answers they received. Moreover, every answer is associated with 12 aspect labels. Using the labels attained through this editorial study, we measure the correlation between the aspects in our taxonomy and the perceived usefulness of answers.

The main contributions of our work are the following:

- Proposing a taxonomy of aspects to capture the usefulness of an answer in the question answering process.
- Conducting an editorial study to create a labeled dataset,[1] to investigate the correlation between various aspects of answers and their perceived usefulness.
- Assessing the effectiveness of four commonly used answer quality measures (ROUGE [11], BLEU [16], METEOR [2], and BERTScore [26]) in capturing the usefulness of answers.

We address the following research questions:

- RQ1 [Constituents of perceived usefulness]: What makes an answer useful in the eyes of the asker? Which aspects play the most important role in capturing usefulness?

CHIIR '21, March 14–19, 2021, Canberra, ACT, Australia

ACM ISBN 978-1-4503-8055-3/21/03…$15.00
https://doi.org/10.1145/3406522.3446028

[1] Dataset is available at https://github.com/barla/AnswerUtilityDataset.

- RQ2 [Impact of attributes]: Is usefulness affected by certain attributes? Do attributes like a question's importance to the asker, or the answerer's interest in the question play a role?
- RQ3 [Human versus system answers]: Are system-generated answers on a par with human answers in terms of usefulness? What makes a human answer different from an answer given by a state-of-the-art question answering system?
- RQ4 [Answer quality measures] Are commonly used answer quality measures able to capture the usefulness of an answer completely? Do they capture all aspects of usefulness or are they biased towards a subset of them?

Some of our selected findings are as follows:

- Relevance, correctness, completeness, and comprehensiveness are the most important four aspects, having relatively high correlation with the perceived usefulness of answers given to non-factoid questions.
- Askers assess the usefulness of answers mainly based on the answer itself, while attributes such as the question's importance and perceived difficulty to the asker, or certain attributes of the answerer do not play an important role.
- Advanced result snippets in commercial web search engines are perceived as superior to human answers, on average. Such snippets are observed to capture the important aspects of usefulness very well.
- Popular answer quality measures have weak correlation with important aspects and, in turn, the perceived usefulness of answers, indicating substantial room for improvement.

The rest of the paper is organized as follows. In Section 2, we present our aspect taxonomy. Section 3 details the conducted editorial study. Our results are presented in Section 4. Section 5 provides a survey of the related work. The paper is concluded in Section 6.

2 ASPECTS OF ANSWER UTILITY

To initiate our understanding of what makes an answer useful, we devised a taxonomy of aspects as potential indicators of usefulness. First, a draft taxonomy was created, involving more than a dozen aspects suggested by the authors and extended by reviewing aspects used in other problem domains, such as online news [1] and community question answering (CQA) [22]. Then, a test study was conducted among the authors to assess the value of each aspect in capturing usefulness. The study involved 30 question/answer pairs generated by the authors. The contribution of each aspect to the perceived answer utility was assessed and certain aspects were removed from the taxonomy. For example, some sentiment-related aspects were removed as they were not observed to have any effect. Aspects about the source of answers, such as reputability and authoritativeness, were also removed since assessing these aspects would require sharing participant details in the actual editorial study. The final taxonomy includes the 12 aspects shown in Figure 1, where the aspects are grouped under four headings. Below, we provide a brief description of each aspect ("The answer is ..."):

(1) relevant as it is about the subject of the question.
(2) correct as it contains an accurate response to the question.
(3) complete as it covers every aspect of the question.
(4) comprehensive as it contains detailed information.
(5) brief as it does not contain redundant information.

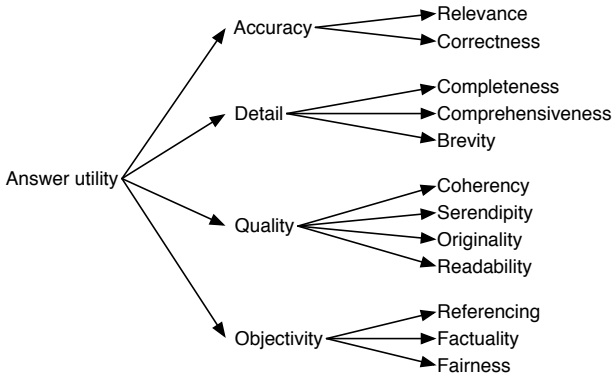


Figure 1: Aspects of answer utility.

(6) coherent as it does not contain inconsistent statements.
(7) serendipitous as it contains some unexpected but positively surprising information.
(8) original as it is not taken from another source.
(9) readable as it is fluently written.
(10) referencing additional information sources.
(11) factual as it is based on things that are known to be true.
(12) fair as it is free of any kind of bias.

We note that, although some of the aspects seem similar at first sight, they are semantically different. For example, comprehensiveness differs from completeness in that an answer may cover every aspect of the question, but to different levels of detail. Similarly, factuality differs from correctness since the correctness of some factual answers may be disputed (e.g., consider the question–answer pair "Do aliens exist?", "Aliens do not exist. No space agency has released evidence about their existence.").

3 EDITORIAL STUDY

The editorial study was conducted with 12 participants, who were hired online through a participant database of a university. Before the study, an information sheet was given to the participants, explaining the details of the study. All participants signed a consent form, declaring that they fully understood the purpose of the study and agreeing that their non-identified data can be used. All communication with the participants was performed online, via email. Assessments were carried out using spreadsheets that the study coordinators provided to the participants. Participants completed pre-study and post-study questionnaires, which collected participant information and feedback, respectively. At the end of the study, each participant was compensated with a gift voucher valued at $500 (AUD). The study was reviewed and approved by the Human Research Ethics Committee of RMIT University.

The study involved three main tasks: (i) question generation and labeling, (ii) answer generation and labeling, and (iii) aspect labeling. In the first task, each participant acted as an asker and generated a number of questions that were used in the subsequent steps of the study. In the second task, each participant acted as an answerer and generated answers for questions received from an asker. Concurrent to this task, additional "system" answers were retrieved from the Google web search engine. In the third task, all answers were returned to the askers of respective questions, and

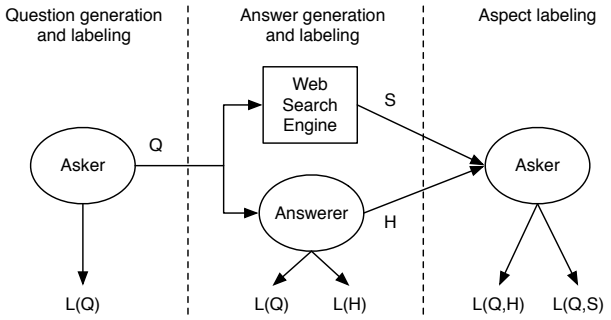


Figure 2: Editorial study workflow: Q is a question; H and S are human and system answers, respectively; and L are label sets obtained using Q, H, and S.

then the askers labeled the various aspects of every answer they received for their questions, both human- and system-generated. The workflow of the editorial study is illustrated in Figure 2.

3.1 Pre-study Questionnaire

The participants were asked to provide information about their age, gender, English fluency, education level, area of expertise, and computer proficiency. Participant age ranged from 18 to 57, with an average of 32. There were more female participants (f=10, m=2), but we do not expect this to affect our results substantially. Ten of the participants were native speakers of English, while two were proficient. Three participants had a master's degree, five had a bachelor's degree, one had a graduate certificate, and three had a high school degree. The expertise areas of the participants were quite varied (marketing, engineering, education, graphic design, communication, linguistics, financial planning, arts, economics, accounting). Ten participants said they use a computer many times everyday, and two said they use a computer once a day.

3.2 Task 1: Question Generation and Labeling

The participants (askers) were first presented with a non-factoid question intent taxonomy containing six categories (how to, cause/effect, description, comparison, advice, and debate). The taxonomy was developed by the authors and validated through crowdsourcing studies.[2] The categories were explained to the askers in detail, together with a sample set of question patterns for each intent category. They were then asked to select four questions from their recent search history, i.e., queries they previously submitted to a web search engine, for each category, as well as four additional factoid questions to enable comparative analysis. Thus, each asker generated 28 questions in total.

The askers were requested to provide well-formed questions that start with a question word and end with a question mark. The scope of the questions was not constrained, but the askers were instructed to refrain from providing private questions that may reveal their identity or contain confidential or sensitive information. They were allowed to convert short, keyword-based queries in their search history into well-formed questions. They were also allowed to create new questions if no suitable question was found in their search history. Of the 336 questions generated, 63.1% came from the askers' search history and the remaining 36.9% were created.

[2]We leave the description of the taxonomy as a future work.

Table 1: Distribution of the labels from askers

	Statistics		Labels		
Criteria	μ	σ^2	0	1	2
Question importance	0.80	0.73	0.38	0.43	0.19
Pre-search difficulty	0.80	0.65	0.33	0.54	0.13
Post-search difficulty	1.60	0.56	0.05	0.30	0.65

Generated questions were reviewed by the study coordinators. About 12.8% of the provided questions were found to have issues: the majority were not in correct intent categories; some were not well-formed; and, one question was a duplicate. Another iteration was performed with the askers to fix or replace those problematic questions. The askers were then requested to assess the following for each of their questions using a three-point scale:

(1) Question importance: How important is/was it for you to get a useful answer to this question: not important (0), somewhat important (1), very important (2)?
(2) Pre-search question difficulty: How likely do you think it is for a human to provide a useful answer to this question without consulting any information source: not likely (0), somewhat likely (1), very likely (2)?
(3) Post-search question difficulty: Consider a human who is not able to provide a useful answer to this question before consulting information sources about the answer. How likely do you think it would be for that human to provide a useful answer after consulting some information sources: not likely (0), somewhat likely (1), very likely (2)?

The difficulty of a question was assessed in two parts as pre-search and post-search question difficulty, where the former aimed to capture the inherent complexity and knowledge requirement of the question, while the latter aimed to capture the availability and accessibility of information sources that can be used to answer the question. During the labeling, askers were not allowed to consult any information source to prevent a bias in assessment of pre-search and post-search difficulty. Throughout the study, an information source was defined as an offline (e.g., a hard-copy book, a real-life friend) or an online (e.g., a web search engine, a community question answering site, a social media application) resource that could be used to seek a useful answer to a given question.

According to Table 1, 19% of the questions were deemed "very important" by their askers, while questions that were deemed "not important" were about two times more common. The low frequency of "very important" questions is potentially due to questions that did not come from askers' search history, but were instead generated artificially. Indeed, when we take the question source into account (not shown in the table), we observe that the number of "very important" and "not important" questions each constitute about one-fourth of askers' search history. On the other hand, "very important" questions are seven times less common than "not important" questions within the set of artificially generated questions.

In Table 1, one-third of the questions are seen (pre-search) as "not likely" to receive a useful answer, and only 13% of the questions are seen as "very likely" to receive a useful answer. The mean value for the estimated pre-search useful answer likelihood is 0.80,

Table 2: Distribution of the labels from answerers

Criteria	Statistics		Labels		
	μ	σ^2	0	1	2
Answerer's interest	0.77	0.71	0.39	0.44	0.16
Answerer's knowledge	0.52	0.70	0.60	0.27	0.12

which indicates that the generated questions tend to be perceived as difficult, on average. The distribution of post-search labels are quite different to pre-search labels. The askers believe that only 5% of the questions are "not likely" to receive a useful answer, while about two-thirds are "very likely" to get a useful answer. This may be a sign of the participants' confidence in the wide-spread availability and easy accessibility of online information.

3.3 Task 2: Answer Generation and Labeling

Each set of 28 questions generated by an asker was provided to another participant (answerer), who was requested to come up with answers to the questions they received. The answerers were allowed to consult any information source they wanted. No lower or upper bound was set on the length of the answers they could provide (e.g., their answers could contain a few words or multiple paragraphs). However, they were told that, when providing their answers, the main objective was to come up with answers that can be useful to the person who asked the respective questions. They were also asked to provide the answers in their own words. They could paraphrase or summarize information they acquired from other information sources or expand it with their own knowledge, but they were not allowed to copy and paste content from online sources. Before answering any question assigned to them, the answerers provided the following information for every question:

(1) Answerer's interest: How interested would you be in getting an answer to this question: not interested (0), somewhat interested (1), very interested (2)?
(2) Answerer's knowledge: How likely are you to give a useful answer to this question without consulting any information source: not likely (0), somewhat likely (1), very likely (2)?

According to Table 2, the answerers were "very interested" in only a small portion of the questions they received (roughly, one out of every six questions). However, they were "moderately interested" or "very interested" in 60% of questions, an encouraging sign of participants' engagement in the study. Regarding their knowledge about the question, the answerers felt that they were "not likely" to provide a useful answer for the majority of the questions (60%), while they claimed that they were "very likely" to provide a useful answer for only 12% of the questions. The latter value is very close to the percentage of questions (13%) that were predicted by the askers as "very likely" to receive a useful answer (see the discussion associated with Table 1 in Section 3.2). However, when the two tables are compared, we also observe that the askers were more optimistic regarding the percentage of questions that were "not likely" to receive a useful answer (33% versus 60%).

After answering their questions, the answerers assessed the usefulness of their answers ("How useful do you think the answer you provided will be for a person who asks the corresponding question?") using a 5-point scale: not useful (0), slightly useful (1), moderately useful (2), useful (3), very useful (4). Table 3 reports the distribution of labels for usefulness estimated by the answerers. We observe the distribution to be relatively balanced with the exception of the "not useful" label, assigned to only 6% of answers.

Table 3: Distribution of the labels assigned by the answerers regarding the estimated usefulness of their answers

	Statistics		Labels				
	μ	σ^2	0	1	2	3	4
Usefulness	2.36	1.18	0.06	0.20	0.29	0.25	0.21

Table 4: Sorted list of answer sources used by answerers

Answer source	Frequency
Search engine result page	0.38
Myself (answerers themselves)	0.27
Other website (excluding CQA and news websites)	0.23
News website	0.04
Community question answering website	0.04
Another person	0.02
Mobile or desktop application	0.01
Offline resource	0.01
Personal assistant	0.00
Other resource (excluding any resource above)	0.00

Finally, the answerers also stated which information source they mainly used when generating each answer (myself/another person/offline resource (e.g., a hard-copy book)/personal assistant (e.g., Alexa)/mobile or desktop application (e.g., Calendar)/search engine result page (e.g., Google)/community question answering website (e.g., Quora)/news website (e.g., CNN)/other website/other resource). They were instructed to select the "search engine result page" option only if the answer was available somewhere on the displayed search engine result page. They were asked not to select this option when the search engine's only role was to facilitate navigation to a website that actually provided the answer.

Table 4 shows the frequency distribution of resources used by the answerers. The great majority of the answers are obtained from three main sources: a search engine result page; the answerers themselves; or, a website excluding CQA and news websites). Interestingly, no participant obtained their answers from a personal assistant, despite their increasing usage in homes and mobile phones. Also, no participant selected the "other resource" option, indicating that the options we have provided in the study had an extensive coverage of resources that can be used in question answering.

Concurrent to answer generation, we obtained a single "system" answer for each of the 336 questions we have. To this end, we submitted questions to Google as queries and extracted answers from three different modules on the retrieved search engine result pages (SERP): featured snippets module (FSM), knowledge-base module (KBM), and search result module (SRM). Example answers extracted from these modules are shown in Figure 3. We considered all search verticals that present information pulled from knowledge bases, structured databases, or internal lexicons (e.g., the dictionary module) as part of KBM. All answers were scraped manually to prevent erroneous query rewriting, occasionally performed by Google.

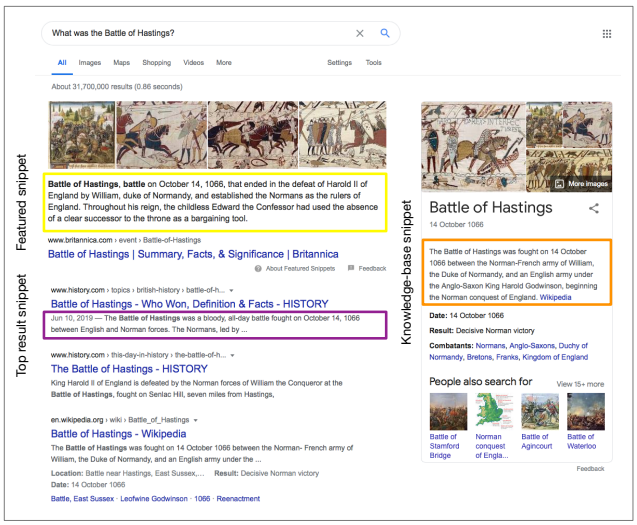


Figure 3: Three types of system answers extracted from Google for the query "What was the Battle of Hastings?".

In the case of SRM, the top result snippet in the module formed our answer. This module was always available on the SERPs we scraped, but it usually provided low-quality answers. Therefore, we extracted the answer from the SRM module as a last resort option, only when the FSM and KBM modules were both unavailable on the SERP. When the FSM and KBM modules were available at the same time (this was rather uncommon), we prioritized FSM over KBM for answer extraction. This way, we obtained a single system answer for each question in our sample. FSM, KBM, and SRM contributed 57%, 11%, and 32% of our system answers, respectively.

3.4 Task 3: Aspect Labeling

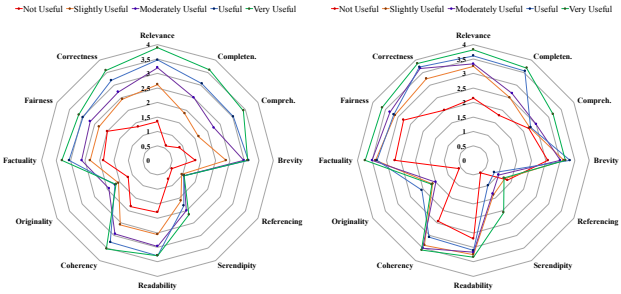
In this final task, the askers were provided with two different answers (a human answer and a system answer) for each of the 28 questions they have previously generated, forming 56 question–answer pairs in total. The askers were then requested to assess the usefulness of the answers they received ("How useful do you think the provided answer is from your point of view?") using a 5-point scale: not useful (0), slightly useful (1), moderately useful (2), useful (3), very useful (4). During the assessment of usefulness, they were allowed to consult any information source, except for Google, since the system answers they received had been retrieved from Google.

Table 5 shows the distribution of the labels assigned by the askers regarding the usefulness of answers they received. According to the mean values reported in the table, the askers seem to find human answers slightly more useful than the system answers. However, as we will see later in Section 4.3, this observation is somewhat misleading since the usefulness of system answers show high variation depending on which module the answer was retrieved from.

Finally, the askers assessed the aspects for each of the answers they received. They were presented with the 12 statements given in Section 2 and declared their agreement using a 5-point scale: strongly disagree (0), disagree (1), neither agree nor disagree (2), agree (3), strongly agree (4). The order of questions was randomized before they were assessed. The participants were suggested to take a short break after completing the labeling of each aspect.

Table 5: Distribution of the perceived usefulness labels assigned by the askers for human and system answers

	Statistics		Labels				
	μ	σ^2	0	1	2	3	4
Human	2.47	1.25	0.10	0.14	0.20	0.33	0.24
System	2.33	1.47	0.18	0.15	0.14	0.25	0.29



(a) Non-factoid questions **(b) Factoid questions**

Figure 4: Mean values of different aspect labels.

3.5 Post-study Questionnaire

The participants filled out a post-study questionnaire to let the study coordinators know about their overall experience with the study (satisfied/neutral/unsatisfied), whether they found the study challenging (easy/medium/difficult), and whether the compensation amount was fair (should be less/fair/should be more). Of the 12 participants, 10 were satisfied with the study and two were neutral. Three participants found the study easy, while seven found it medium difficulty and two found it difficult. Finally, the great majority of the participants found the compensation amount fair, while two participants felt that the amount should have been less.

4 RESULTS

Each of the following four sections addresses one of the research questions raised in Section 1. For correlation analysis, Spearman's rank correlation [13] is used since the data is ordinal. The result tables indicate statistical significance for a standard test of the null hypothesis that the correlation is zero. We apply the Bonferroni correction to the p-values to account for multiple hypothesis tests. When we compare correlation coefficients with each other, we follow the method described by Myers and Sirois [14], who use Fisher's z-transformation for correlation coefficients and then test the null hypothesis that $\rho_1 - \rho_2 = 0$. When referring to the strength of the correlation values reported in the tables, we adopt the scale and terminology used by Prion and Haerling [17]: $[0.00, 0.20)$ (negligible), $[0.20, 0.40)$ (weak), $[0.40, 0.60)$ (moderate), $[0.60, 0.80)$ (strong), and $[0.80, 1.00]$ (very strong). The labels obtained from the answerers and askers regarding the usefulness of answers are referred to as the estimated and perceived usefulness, respectively.

4.1 RQ1: Constituents of Perceived Usefulness

To analyze the impact of various aspects on the perceived usefulness of answers, we display the mean values of each aspect for different levels of usefulness, using radar graphs in Figure 4. We also compute the correlation values between the aspects and perceived usefulness.

Table 6: Correlations between the perceived usefulness and aspects of answers (non-factoid questions)

	Aspects												Perc.
	Compl.	Rele.	Compr.	Corr.	Cohe.	Fact.	Read.	Brev.	Fair.	Sere.	Refe.	Orig.	usef.
Completeness		**0.67****	**0.66****	**0.65****	0.52**	0.52**	0.42**	0.50**	0.41**	0.27**	0.09	0.11	**0.65****
Relevance			0.49**	0.59**	0.5**	0.44**	0.35**	0.47**	0.34**	0.32**	−0.02	0.15*	**0.65****
Comprehens.				0.55**	0.45**	0.46**	0.40**	0.34**	0.41**	0.29**	0.21**	0.01	**0.63****
Correctness					0.55**	0.58**	0.40**	0.45**	0.47**	0.26**	−0.04	0.01	**0.60****
Coherency						0.46**	**0.62****	0.52**	0.43**	0.33**	0.03	0.13	0.51**
Factuality							0.43**	0.33**	0.57**	0.13	0.02	−0.03	0.47**
Readability								0.39**	0.39**	0.23**	0.04	0.13	0.37**
Brevity									0.38**	0.13	0.10	0.14*	0.37**
Fairness										0.14*	−0.07	−0.07	0.33**
Serendipity											0.07	0.14	0.26**
Referencing												0.33**	0.08
Originality													0.08

* *significance level* $p < 0.000641$, ** *significance level* $p < 1.282\text{e}{-05}$. *(Bonferroni corrected from* $p = 0.05$ *and* $p = 0.001$*, 78 tests)*

Table 7: Correlations between the perceived usefulness and aspects of answers (factoid questions)

	Aspects												Perc.
	Rele.	Corr.	Compl.	Fact.	Sere.	Cohe.	Fair.	Compr.	Orig.	Read.	Brev.	Refe.	usef.
Relevance		**0.65****	**0.69****	0.44**	0.18	0.42*	0.45**	0.28	0.27	0.32	0.47**	−0.16	0.55**
Correctness			**0.61****	0.48**	0.18	0.47**	0.44**	0.25	0.20	0.35*	0.41*	−0.16	0.53**
Completeness				0.42*	0.18	0.45**	0.48**	0.46**	0.08	0.36*	0.26	−0.19	0.52**
Factuality					0.11	0.56**	0.57**	0.46**	0.06	0.43**	0.43*	−0.01	0.43*
Serendipity						0.11	0.04	0.32	0.23	0.03	−0.18	0.27	0.36*
Coherency							0.55**	0.29	0.06	**0.73****	0.42*	−0.23	0.34
Fairness								0.20	0.20	0.39*	0.55**	−0.19	0.31
Comprehens.									−0.16	0.24	0.03	0.23	0.28
Originality										−0.14	0.20	0.25	0.23
Readability											0.30	−0.16	0.17
Brevity												−0.08	0.13
Referencing													−0.01

* *significance level* $p < 0.000641$, ** *significance level* $p < 1.282\text{e}{-05}$. *(Bonferroni corrected from* $p = 0.05$ *and* $p = 0.001$*, 78 tests)*

The computed values are reported in the right-most columns of Tables 6 and 7 for non-factoid and factoid questions, respectively, together with the correlation values between pairs of aspects. When discussing the results in this section, we focus only on statistically significant correlations and their significant differences (Bonferroni corrected equivalent of $\alpha = 0.05$ and $\alpha = 0.001$, see table captions).

For non-factoid questions, relevance, completeness, correctness, and comprehensiveness have all strong correlation with the perceived usefulness, while coherency and factuality have moderate correlation with it. Besides, the correlations of these four aspects with usefulness are statistically significantly higher ($p < 0.00076$, Bonferroni corrected from $p = 0.05$, 66 tests) than the correlation between any other aspect, except for coherency and factuality. Among the 66 possible pairs of aspects, only four pairs are strongly correlated, indicating that our taxonomy provides a diverse set of aspects that are semantically different. Interestingly, the three of those four pairs involve aspects that are strongly correlated with usefulness. The moderate correlation between correctness and factuality justifies our claim (see Section 2) that these two aspects are not identical. Finally, the mean values displayed in Figure 4a indicate that answers to non-factoid questions are generally perceived as useful even when they are not original and not supported by references, as long as some influential aspects are present. In the same figure, we also observe that "not useful" answers are more likely to be incomplete or less detailed than being irrelevant or incorrect.

For factoid questions, none of the aspects have a strong correlation with usefulness, while relevance, correctness, completeness, and factuality are the most prominent aspects having moderate correlation with perceived usefulness. Comprehensiveness, which is an important aspect for non-factoid questions, is absent in this list, probably because most factoid questions are likely to be answered satisfactorily with a phrase or a short sentence. In our data, "useful" and "very useful" answers have a mean length of 134.1 and 252.2 characters for factoid and non-factoid questions, respectively (Student's t-test, $t = -5.68$, $p < 0.001$), supporting this claim.

4.2 RQ2: Impact of Attributes

We next investigate whether certain attributes (importance and pre-search/post-search difficulty of the question as judged by the asker; answerer's interest and knowledge about the question; or, answer usefulness estimated by the answerer) play a role in the perceived usefulness of answers. Intuitively, one may expect that

Table 8: Correlation of certain attributes with the perceived usefulness of answers and their aspects

	Aspects												Usefulness	
	Rele.	Corr.	Compl.	Compr.	Brev.	Refe.	Sere.	Read.	Cohe.	Orig.	Fact.	Fair.	Est.	Perc.
Importance	−0.11	−0.17	−0.14	0.00	−0.04	0.02	−0.05	−0.01	−0.06	0.00	−0.07	0.01	0.05	−0.05
Pre-search diff.	0.12	0.10	0.10	0.12	0.06	0.02	0.01	0.02	0.11	0.06	0.05	0.05	0.04	0.06
Post-search diff.	0.08	0.33**	0.13	0.11	0.19*	−0.12	−0.20*	0.01	0.12	−0.18	0.24**	0.31**	0.24**	0.14
Interest	−0.06	0.00	0.05	0.12	−0.12	−0.05	−0.03	−0.05	−0.13	−0.06	0.02	−0.04	0.18	−0.02
Knowledge	0.02	−0.02	0.11	0.20*	−0.05	0.12	−0.1	−0.03	−0.08	0.01	0.09	0.01	0.04	0.06
Est. usefulness	0.27**	0.29**	0.30**	0.12	0.12	−0.22*	−0.04	0.14	0.13	−0.18	0.32**	0.33**	**1.00****	0.16

* *significance level* $p < 0.000595$, ** *significance level* $p < 1.19\text{e}{-05}$. *(Bonferroni corrected from* $p = 0.05$ *and* $p = 0.001$*, 84 tests)*

Table 9: Various statistics about the perceived usefulness of human and system answers

Que. type	Ans. type	Human		System		#	$\mu_h - \mu_s$
		μ_h	σ_h^2	μ_s	σ_s^2		
All	FSM	2.43	1.25	**2.88**	1.19	192	−0.44*
	KBM	2.53	1.41	**3.08**	1.15	38	−0.55
	SRM	**2.53**	1.21	1.07	1.21	106	1.46*
	All	**2.47**	1.25	2.33	1.47	336	0.15
F	FSM	2.38	1.47	**3.05**	1.28	21	−0.67
	KBM	2.74	1.63	**3.53**	1.07	19	−0.79
	SRM	**3.38**	1.41	1.75	1.98	8	1.63
	All	2.69	1.53	**3.02**†	1.45	48	−0.33
NF	FSM	2.44	1.23	**2.85**	1.18	171	−0.42*
	KBM	2.32	1.16	**2.63**	1.07	19	−0.32
	SRM	**2.46**	1.17	1.01	1.13	98	1.45*
	All	**2.44**	1.20	2.21†	1.44	288	0.23

* *sig. diff. btw. human/system* $p < 0.05$ *(Tukey HSD).*
† *sig. diff. btw. factoid/non-factoid* $p < 0.05$ *(Tukey HSD).*

less difficult questions or questions which the answerer is interested in or has knowledge about are easier to answer, and thus, they should receive more useful answers. However, when we look at the correlation values reported in Table 8, we observe all attributes to have negligible correlation with perceived usefulness (although none of the reported results are statistically significant). This is probably due to the answerer's ability to access a wide range of information sources when answering the questions they received. As an example, an answerer with little knowledge about a difficult question may still come up with a useful answer to the question after checking various information sources (112 out of 203 such questions were found useful). It is also striking to see that the usefulness of answers estimated by the answerers (the last row) and that perceived by the askers have negligible correlation. This implies that the answerers often could not accurately identify what a useful answer would look like in the eyes of the asker.

4.3 RQ3: Human Versus System Answers

We compare human and system answers for two different sets of questions (factoid and non-factoid). We further divide each of these two sets into three groups, depending on which module on Google's SERP the corresponding answer was extracted from: featured snippet module (FSM); knowledge-base module (KBM); and, search result module (SRM). In Table 9, we report the mean and standard deviation for the perceived usefulness of human and system answers, separately, for each question group. We also show the difference between the mean values for perceived usefulness of human and system answers in the last column.

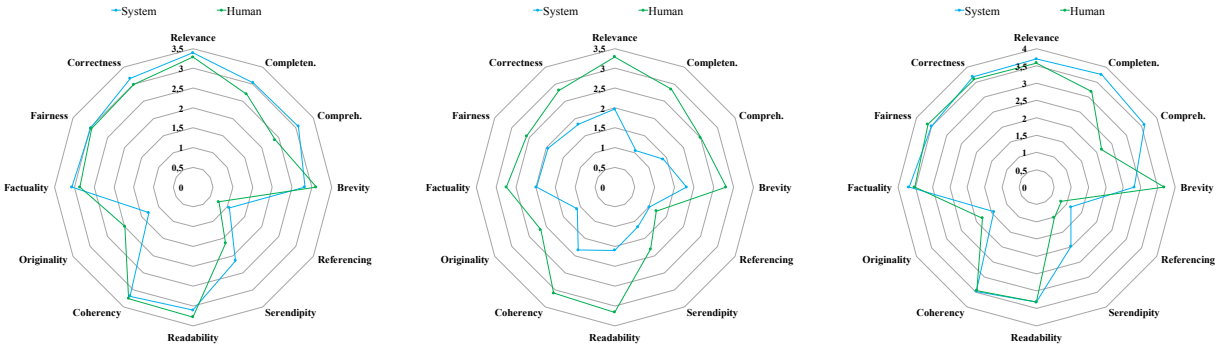
In Table 9, the reported differences for KBM answers do not show statistical significance for factoid or non-factoid questions, potentially due to small sample sizes (reported in the seventh column of the table). Hence, we do not analyze and discuss these results further. Focusing on non-factoid questions only, we observe that the FSM answers are perceived as more useful than human answers on average (the difference of means is -0.42). On the contrary, the SRM answers are perceived as somewhat less useful than human answers (the difference of means is 1.45). When we compare the mean values for factoid and non-factoid question sets, we see that the system answers given to factoid questions are perceived as statistically significantly more useful (3.02 versus 2.21), while there is no statistically significant difference in the case of human answers (2.69 versus 2.44). This may imply that modern search engines are effective in answering factoid questions, but non-factoid questions have room for improvement.

Figure 5 compares the mean values of aspects for human and system answers. Since the KBM answers have a small sample size and have similar quality to the FSM answers, we combine the two sets in our analysis (Figures 5a and 5b). We display the mean values for SRM answers only for non-factoid questions (Figure 5c) since we have just eight factoid questions with an SRM answer.

According to Figure 5a, the FSM/KBM answers given to non-factoid questions are seen as relatively more complete, comprehensive, and serendipitous than human answers. However, human answers are perceived as more original (recall that the answerers had been instructed to avoid copying text from online sources). According to Figure 5b, the mean values associated with the SRM answers are lower than those of human answers, for all aspects. The largest difference is observed in the completeness and readability aspects. This is manly because most search result snippets are created by concatenating partial sentences extracted from different parts of a web page. Finally, according to Figure 5c, the main difference between the human answers and FSM/KBM answers given to factoid questions is in the amount of provided information. While human answers are brief, FSM/KBM answers are more comprehensive and serendipitous. Also, it appears that it was more difficult for participants to distinguish the originality of the answers.

4.4 RQ4: Answer Quality Measures

ROUGE [11], BLEU [16], METEOR [2], and BERTScore [26] are some commonly used measures for evaluating the quality of question answering systems. These measures yield a quality score for a



(a) Non-factoid, FSM∪KBM. **(b) Non-factoid, SRM.** **(c) Factoid, FSM∪KBM.**

Figure 5: Mean values of aspect labels for human and system answers.

Table 10: Correlations of quality measures with the perceived usefulness of answers and their aspects

	Aspects												Perc.
	Rele.	Corr.	Compl.	Compr.	Brev.	Refe.	Sere.	Read.	Cohe.	Orig.	Fact.	Fair.	usef.
ROUGE-L	0.36**	0.41**	0.31**	0.27*	0.35**	−0.11	0.19	0.25*	0.29*	−0.03	0.28*	0.23*	0.34**
ROUGE-1	**0.37****	0.42**	0.31**	0.30**	0.33**	−0.12	0.22	0.26*	0.32**	−0.08	0.26*	0.23*	0.39**
ROUGE-2	0.33**	**0.44****	0.28*	0.28*	0.35**	−0.06	0.19	0.22	0.28*	−0.08	0.28*	0.25*	0.31**
BLEU	0.36**	0.43**	0.33**	**0.37****	0.26*	−0.07	**0.27***	0.19	0.23*	−0.16	0.29**	0.24*	0.38**
METEOR	0.35**	0.40**	0.28*	0.23*	**0.40****	−0.09	0.15	**0.27***	0.31**	0.03	0.25*	**0.28***	0.34**
BERTScore	0.33**	0.42**	**0.36****	0.32**	0.26*	−0.13	0.19	0.25*	**0.33****	−0.06	**0.31****	0.22	**0.42****

* *significance level* $p < 0.000641$, ** *significance level* $p < 1.282\text{e}{-05}$. *(Bonferroni corrected from* $p = 0.05$ *and* $p = 0.001$*, 78 tests)*

given question, candidate answer pair by comparing the candidate answer with a reference answer which is assumed to be a perfect answer for the question. Herein, we aim to understand the effectiveness of these measures in capturing the perceived usefulness of answers, specifically focusing on non-factoid questions.

To obtain a sample set of questions with reference and candidate answers, we adopt the following process: We first remove questions with no answers labeled as "useful" or "very useful", leaving 216 questions (out of the 288 questions available). For each of the remaining questions, we select one of the two answers (a human and a system answer) associated with the question as a reference answer. Here, we prefer "very useful" answers over "useful" answers. The ties (about 20% of the remaining questions) are broken by selecting the answer with the higher mean value computed over all aspects. Of the 216 reference answers obtained, 116 are human answers, and the remaining 100 are system answers. The answers that are not selected in the process are used as candidate answers. The reference answers are statistically significantly more useful than the candidate answers, on average: 3.51 versus 1.92 (Student's t-test, $t = 17.6$, $p < 0.01$).

We compute the four measures above using our reference and candidate answers. Table 10 shows the correlation between the computed measures and the perceived usefulness of answers. According to the table, all measures exhibit statistically significant, but weak to moderate correlation with perceived usefulness, BERTScore performing slightly better than the rest. Also, the measures are observed to have relatively high correlation with the important aspects identified previously (relevance, correctness, completeness, and comprehensiveness). Finally, no measure's correlation with an aspect is statistically significantly different from the correlation of another measure with the same aspect.

5 RELATED WORK

Shah and Pomerantz [22] conducted a crowdsourcing study to evaluate the quality of answers in CQA. They measured the correlation between answers' quality and 13 aspects, previously proposed by Zhu et al. [28] (informative, polite, complete, readable, relevant, brief, convincing, detailed, original, objective, novel, helpful, expert). The study was somewhat inconclusive and reported no tangible correlation between answer quality and the considered aspects. Furthermore, a classifier was trained using the 13 aspects as features in order to predict the best answers. The classifier was reported to yield worse results than a naive majority class predictor.

Our work differs from the work mentioned above in several ways. First, we focus on non-factoid question answering, instead of CQA, i.e., the problem domains are not the same. Second, we conduct an editorial study instead of a crowdsourcing study. This lets us work closely with the participants and train them better for the tasks at hand, in contrast to a crowdsourcing study involving loosely coordinated and poorly trained workers. Third, the participants in the setup of Shah and Pomerantz assess answers provided to questions asked by anonymous people, while the participants in our study assess answers given to their own questions, rendering the study more realistic. Fourth, although there is a certain overlap between the aspects used in the two works, our dependent variable is the usefulness of answers, which is different from answer quality, a relatively vague target. Finally, we report high correlation between certain aspects and answer utility, unlike the work of Shah and Pomerantz, reporting inconclusive results.

Another editorial study was conducted by Arapakis et al. [1] to assess the quality of online news articles. They used 14 different aspects to model the quality of news articles (fluency, conciseness, descriptiveness, novelty, completeness, referencing, formality, richness, attractiveness, technicality, popularity, subjectivity, sentimentality, polarity). Their problem domain as well as the considered aspects differ from ours since there is no notion of questions and answers in their problem context.

While some studies aim to understand what makes an answer high quality, most of these studies focus on a limited set of aspects and are less comprehensive than ours. Lin et al. [12] investigated the optimal verbosity of answers (e.g., phrase, sentence, paragraph, document) through a user study. Users were shown to prefer paragraph-size answers, while the reliability of the answer source and the size of the search task were not found to have a significant effect on the optimal answer size. Hart and Sarma [6] conducted a crowdsourcing study to investigate the impact of an answerer's social reputation and the verbosity of answers on the perceived answer quality in CQA. They found that novice users are likely to judge the quality of answers mainly based on some intrinsic features of the answers, such as presentation and content, instead of relying on social cues. Also, users are more interested in factors such as thoroughness and conciseness instead of answer length. Lee et al. [10] suggested that polite answers are more likely to be perceived as high quality, pointing at a politeness bias. Fichman [4] compared the answer quality of four different CQA websites and found that retrieving answers from more sources yields more complete and verifiable answers, but does not result in more accurate answers. Qu et al. [18] conducted a crowdsourcing study to observe the interaction of users with answers retrieved from a non-factoid question answering system, showing that users react to good and bad answers somewhat differently, and they can identify good answers relatively fast.

Several works model high-quality answers using low-level features extracted from questions, answers, and other sources. Fu et al. [5] trained a model using 24 textual and non-textual features to predict the quality of answers in CQA, and found that review and user features are the most powerful indicators of a high-quality answer, while the usefulness of content features vary depending on the knowledge domain. Le et al. [9] used four different groups of features (personal, community-based, textual, and contextual) to train a model to determine what constitutes the quality of answers given in the education domain. Shah [21] predicted best answers using a model trained with features extracted from the interaction history of askers and answerers. Yao et al. [25] focused on early detection of high-quality question-answer pairs in CQA. Kucuktunc et al. [8] showed that the best answers in the business domain tend to be more neutral while those in the news domain are more positive in terms of the sentiments expressed in the answers. Hashemi et al. [7] designed a neural network architecture to predict the quality of answers in non-factoid question answering systems.

Finally, a concurrent line of research exists in the context of relevance modeling in information retrieval systems. These works investigate the meaning of relevance and the aspects contributing to it. Interested readers are referred to the works of Borlund [3], Xu and Chen [24], Zhang et al. [27], and Saracevic [20].

6 CONCLUSION

This work investigated the impact of various aspects on the perceived usefulness of answers in non-factoid question answering. We found that the usefulness of answers is strongly correlated with four aspects: relevance, correctness, completeness, and comprehensiveness. That is, when assessing usefulness, users care more about the accuracy and detail, and less about the quality and objectivity of answers. We note that the reported values are aggregates over six non-factoid question intent categories. The importance of aspects may differ if we focus on a particular intent category (e.g., objectivity may become more prominent for "advice" questions). We defer intent-level analysis to future work due to a lack of space.

There is negligible correlation between the perceived usefulness of answers and certain attributes (e.g., the importance and difficulty of the question according to the asker, or the answerer's interest and knowledge about the question). However, we believe that this finding is interesting as it implies that the perceived usefulness is much more about the answer itself, rather than the context surrounding it. In this respect, our finding is in line with the work of Hart and Sarma [6], who showed the limited role of social cues.

As another contribution, we provided a comparison of the perceived usefulness of human and system answers. Our results showed that system answers retrieved from Google's featured snippet and knowledge-base modules were perceived as more useful than even human answers. This confirms the recent advances in question answering systems and is in line with various online challenges, where novel deep learning models are reported to outperform humans [15, 19, 23]. Finally, we found that BERTScore slightly outperforms three commonly used answer quality measures (ROUGE, BLEU, and METEOR) in capturing the usefulness of answers. In general, all four measures were observed to have weak correlation with the most important aspects of usefulness, indicating potential room for improvement.

Our work has several implications for question answering and search systems in practice. First, while most existing systems focus on answer relevance and correctness, we show that other aspects such as completeness and comprehensiveness are almost equally important. Although challenging, deciding the optimal verbosity of answers displayed to users may have practical importance, especially for commercial web search engines, where multiple modules (verticals, answers, search results, ads) compete for the limited space on the SERP. Second, the weak yet surprisingly high correlation between the usefulness and serendipity of answers given to factoid questions hints at a potential improvement through the extension of such answers with additional knowledge. Finally, our work illustrates the striking difference in the perceived usefulness of regular result snippets displayed in web search results and more advanced types of snippets, such as featured snippets and knowledge-base snippets, commonly found on Google's SERPs. This raises interesting questions about the trade-off between the quality and cost of snippet extraction algorithms used in practical search systems.

ACKNOWLEDGMENTS

We heartily thank the participants of our editorial study for helping us and making this research possible. This research was supported in part by the Australian Research Council (DP180102687).

REFERENCES

[1] Ioannis Arapakis, Filipa Peleja, Barla Berkant, and Joao Magalhaes. 2016. Linguistic Benchmarks of Online News Article Quality. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1893–1902.

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Association for Computational Linguistics, Ann Arbor, Michigan, 65–72.

[3] Pia Borlund. 2003. The concept of relevance in IR. *J. Assoc. Inf. Sci. Technol.* 54, 10 (2003), 913–925.

[4] Pnina Fichman. 2011. A Comparative Assessment of Answer Quality on Four Question Answering Sites. *J. Inf. Sci.* 37, 5 (2011), 476–486.

[5] Hengyi Fu, Shuheng Wu, and Sanghee Oh. 2015. Evaluating Answer Quality across Knowledge Domains: Using Textual and Non-Textual Features in Social Q&A. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*. American Society for Information Science, USA, Article Article 88, 5 pages.

[6] Kerry Hart and Anita Sarma. 2014. Perceptions of Answer Quality in an Online Technical Question and Answer Forum. In *Proceedings of the 7th International Workshop on Cooperative and Human Aspects of Software Engineering*. Association for Computing Machinery, New York, NY, USA, 103–106.

[7] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2019. Performance Prediction for Non-Factoid Question Answering. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, New York, NY, USA, 55–58.

[8] Onur Kucuktunc, B. Barla Cambazoglu, Ingmar Weber, and Hakan Ferhatosmanoglu. 2012. A Large-Scale Sentiment Analysis for Yahoo! Answers. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, New York, NY, USA, 633–642.

[9] Long T. Le, Chirag Shah, and Erik Choi. 2016. Evaluating the Quality of Educational Answers in Community Question-Answering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. ACM, New York, NY, USA, 129–138.

[10] Shun-Yang Lee, Huaxia Rui, and Andrew B. Whinston. 2019. Is Best Answer Really the Best Answer? The Politeness Bias. *MIS Q.* 43, 2 (2019), 579–600.

[11] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81.

[12] Jimmy Lin, Dennis Quan, Vineet Sinha, Karun Bakshi, David Huynh, Boris Katz, and David Karger. 2003. What Makes a Good Answer? The Role of Context in Question Answering. In *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction*.

[13] Jerome L. Myers, Arnold D. Well, and Robert F. Lorch. 2010. *Research Design and Statistical Analysis*. Routledge.

[14] Leann Myers and Maria J. Sirois. 2006. Differences between Spearman Correlation Coefficients. In *Encyclopedia of Statistical Sciences*. John Wiley and Sons.

[15] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.), Vol. 1773. CEUR-WS.org.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, USA, 311–318.

[17] Susan Prion and Katie Haerling. 2014. Making Sense of Methods and Measurement: Spearman-Rho Ranked-Order Correlation Coefficient. *Clinical Simulation in Nursing* 10 (2014), 535–536.

[18] Chen Qu, Liu Yang, W. Bruce Croft, Falk Scholer, and Yongfeng Zhang. 2019. Answer Interaction in Non-Factoid Question Answering Systems. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. ACM, New York, NY, USA, 249–253.

[19] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 784–789.

[20] Tefko Saracevic. 2016. *The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really?* Morgan & Claypool Publishers.

[21] Chirag Shah. 2015. Building a Parsimonious Model for Identifying Best Answers Using Interaction History in Community Q&A. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*. American Society for Information Science, USA, Article 51.

[22] Chirag Shah and Jefferey Pomerantz. 2010. Evaluating and Predicting Answer Quality in Community QA. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 411–418.

[23] Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. TWEETQA: A Social Media Focused Question Answering Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5020–5031.

[24] Yunjie (Calvin) Xu and Zhiwei Chen. 2006. Relevance Judgment: What Do Information Users Consider beyond Topicality? *J. Am. Soc. Inf. Sci. Technol.* 57, 7 (2006), 961–973.

[25] Yuan Yao, Hanghang Tong, Tao Xie, Leman Akoglu, Feng Xu, and Jian Lu. 2015. Detecting high-quality posts in community question answering sites. *Information Sciences* 302 (2015), 70–82.

[26] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

[27] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. 2014. Multidimensional Relevance Modeling via Psychometrics and Crowdsourcing. Association for Computing Machinery, New York, NY, USA, 435–444.

[28] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2009. A Multi-Dimensional Model for Assessing the Quality of Answers in Social Q&A Sites. In *Proceedings of the 14th International Conference on Information Quality*, Paul L. Bowen, Ahmed K. Elmagarmid, Hubert Österle, and Kai-Uwe Sattler (Eds.). HPI/MIT, 264–265.