Text Cohesion in CQA - Does it Impact Rating?

by

LALIT Mohan Mohan, Jahfar Ali, Syed Mohd Ali Rizwi, Y.Raghu Babu Reddy, Dipti M Sharma

in

The Seventh International Conference on Mining Intelligence and Knowledge Exploration (MIKE 2019)

National Institute of Technology (NIT), Goa, India

Report No: IIIT/TR/2019/-1

Centre for Software Engineering Research Lab
International Institute of Information Technology
Hyderabad - 500 032, INDIA
December 2019

Text Cohesion in CQA - Does it Impact Rating?

Lalit Mohan Sanagavarapu, Jahfar Ali Pichen, Syed Mohd Ali Rizwi, Y. Raghu Reddy, and Dipti Sharma

IIIT Hyderabad, India
{lalit.mohan,jahfar.ali}@research.iiit.ac.in
syedmohdali.rizwi@students.iiit.ac.in
{raghu.reddy,dipti}@iiit.ac.in

Abstract. Community Question and Answer (CQA) platforms are expected to provide relevant content that is not readily available through search engines. With an increase in the number of users and growth of internet, CQA platforms have transitioned from generic to domain specific systems. Expert rating, machine learning and statistical methods are being used for assessing the quality of answers. However, the research on importance of consistency as a quality parameter in the form of text cohesion in CQAs is limited. We extracted 109,113 CQAs from StackExchange related to Information Security of the last 8 years to evaluate text cohesion in answers. An empirical study conducted with 246 participants (Information Security Experts, Software Engineers and Computational Linguists) on the extracted answers stated that lack of text cohesion impacts the rating of answers in CQA. Software Engineers are seekers and viewers of answers, they responded to a survey that lack of text cohesion leads to difficulty in reading and remembering. Information Security Experts providing answers to CQA stated that they need text cohesion for understandability.

Keywords: Crowdsourcing · Question and Answers · Quality · Text Cohesion

1 Introduction

With improved digital literacy, affordable computing devices and growing internet user base (4+ Billion users), crowdsourcing platforms are becoming part of mainstream work. The contributions in macro (logo design, task or procedure oriented questions, software development and others) tasks is growing along with micro (image tagging, language translation, survey, and others) tasks on crowdsourcing platforms. Amazon MTurk, StackExchange, Reddit, Google Maps and Wikipedia are among the popular crowdsourcing systems. Crowdsourcing 'Question and Answers' (Q&A) is commonly referred as Community Question and Answering (CQA); StackExchange and Quora are some the widely used CQA systems. While the motivation to contribute on crowdsourcing platforms can be extrinsic or intrinsic, motivation to contribute on CQA systems is mostly intrinsic (internal satisfaction). With increasing

contributions and varying motivation, the quality of answers are not consistent and considered to contain casual and misleading answers [15] [5]. Machine learning, follow graphs, game theory, rating and other techniques have been the research focus for quality assessment [1]. Most of the research on CQA is to assess the quality of questions, estimate the number of answers to a question, assignment of multiple questions to a single responder, on question type (factoid, procedural and others) based response rate, clicks, syntactic and semantic factors [9] [14]. The quality of answers is our focus area with increasing contributors and Machine2Machine communication.

Quality is a key attribute in software engineering as well; code, documents and other deliverables are assessed for quality. To improve the quality of deliverable, the importance of completeness, consistency and correctness in information capturing and reporting are emphasized and used as software engineering quality principles [3] and ISO/IEC 25010:2011. We hypothesize that *'answers that are complete, consistent and correct are expected to have better ratings and views on CQA platforms'*. Completeness means that the answer to the question contains all the available information or is self-contained. Consistency in an answer indicates that semantic or syntactic terms in the sentence(s) do not contradict with each other. Correctness relates to conformance of the answer to a question with reference to ground truth. The fuzziness/interplay among these quality principles has to be dealt before measuring them independently [23]. Correctness and completeness are with reference to world knowledge and ground truth whereas consistency is related to continuity of topic being discussed that helps better comprehension or interpret-ability. Building ground truth of world knowledge is difficult and never-ending effort [17].

In the current study, we limited our study to the role of consistency in answer text and its impact on ratings and views. Consistency is realised through the interpretability of text, which is composed of 'word interpretation', 'syntactic parsing', and 'semantic integration' [12] that leads to successful formation of a mental representation. Consistency reminds the major property of the text to facilitate formation of the right mental representation. Assessing consistency as text interpret-ability depends on textual and subject variables [16]. The subject variables (coherence) are about relevant knowledge and skill set that the reader brings to the text. Hence, the subject variables are confined to the knowledge of readers [13] within the discourse context of the text. The textual variables are linguistic signals (cohesive devices) that mark the relations between the sequence of text constituents. The textual continuity and the relations between the sequences of textual constituents cannot be determined by a finite number of explicit linguistic markers [6] and are computationally elusive task. It also depends on the cognitive ability of the reader to interpret and construct the continuity in the discourse context. In our study, we included various textual cohesion variables such as adding extraneous information, identification of anaphoric references, and supplying background information as identified by Crossley [7], McNamara [16] and Kintsch [4].

In the experiment, we validated the need of text cohesion on ratings and views of StackExchange[1] related to Information Security domain or topic specific answers. Though there were over 149 features [11] to feature cohesion, we experimented with *sentence linking, order, opposition, reason and purpose* features after dimensionality reduction on a sample of 343 responses. Based on the identified features, a survey was conducted to assess the importance of text cohesion in CQA with participation from 3 different groups - Computational Linguists, Software Engineers and Information Security Experts. The survey results on text cohesion were analyzed to understand the differences in responses across the three groups. In the remainder sections of the paper, we describe in (ii) Section 2, the approach to obtain StackExchange CQA and the process followed for obtaining survey responses; (iii) Section 3, results and analysis on the participants' responses to the survey; (iv) Section 4, conclusion and the potential future work.

2 Approach

We adopted an experimental approach with quantitative analysis on a sample to validate our hypothesis that *'answers that are consistent are expected to have better ratings and views on CQA platforms'*. The steps in our approach for measuring importance of cohesion in CQA are shown in figure 1. We

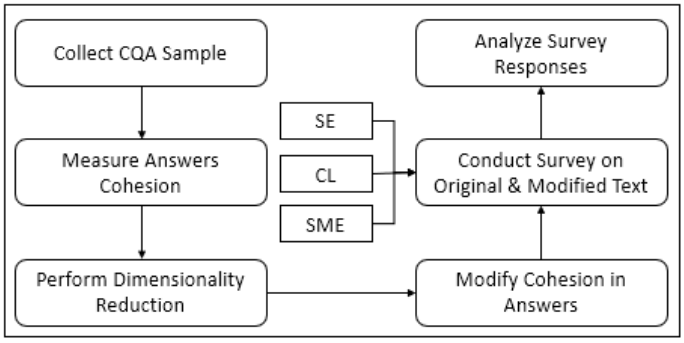

Fig. 1. Process Diagram for Measuring Text Cohesion

limited cohesion assessment to a domain so that a focused group or subject matter experts can be identified for the study. We used StackExchange dataset as it provides an SQL interface to obtain domain specific QA [20]. Using StackExchange SQL interface, we extracted records containing questions, their related answers, question posted date, answer date and the score (user rating) given to the answer using *Question, Answer and Posts* tables into a *.csv* file.

[1] https://security.stackexchange.com/

We selected 'Information Security' as a domain on StackExchange. Attention towards Information security[2] increased with increasing digital usage. As part of the approach, answers rating/score, difference in answer date vis-a-vis question date, cascading/full answers, answer length, count of answers to a question are factored to identify a sample of answers for cohesion analysis. We extracted 109,113 (N) Question and Answers on 'Information Security' for a period of 8 years (StackExchange started in 2009 with very few records and our study happened in 2018, hence, our extract contains data from 2010-17). A stratified sampling is performed on the extracted CQA to identify a sample (S) answers that represents the population N as mentioned in the following steps -

- The length of answers ranged from 38 to 27,596 characters. Approximately 80% of answers have 500 characters. Based on the character size filter, the sample size is reduced to 87,293 (S') records.
- Interestingly, some questions were answered even after 6 years. However, ¿80% of the questions were answered within 15 minutes. With this filter, sample size is further reduced to 74,647 (S") unique answers.
- There are answers that were negatively scored and some answers have been highly scored. As some questions had more than one answer, we selected answers that had the highest score amongst the available answers for a question. The median score or rating of answers is 12 of the initial extract (N) as well as the filtered sample (S"). After the filtering based on median score, the sample size reduced to 343 (S) answers, this sample was used for measuring text cohesion.

To measure cohesion features in text, CohMetrix of McNamara [16] and his co-author Kristopher Kyle's TAACO [8] are widely used tools. We used TAACO (149 features, F) as it has more features (CohMetrix has 106 features) and ease of set-up for identifying various cohesion features in StackExchange answers. To identify impactful, tractable and unrelated features, we performed dimensionality reduction on the cohesion feature values of S answers. While there are many dimensionality reduction techniques [21], we use PCA and Pearson Correlation Analysis for feature identification. On performing PCA, we observed that we can represent all the F features with 90% accuracy by using these 30 features as shown in figure 2. However, PCA gives only a tentative number of features to be used but doesn't give any specific information about the significant features as the values are combined and transformed to principal components. We performed Pearson correlation [19] analysis that examines the relationship between two sets of variables/features. A +1 is a case of perfect direct (increasing) linear relationship (correlation), -1 is a case of a perfect decreasing (inverse) linear relationship (anticorrelation), and some value in the open interval (-1, 1) in all other cases, indicating the degree of linear dependence between the variables. The correlation coefficients of F features of S is available in TextCohesion-Correlation Analysis spreadsheet on Google drive [18]. The features were correlated in the mean

[2] https://tinyurl.com/PWCSecurity

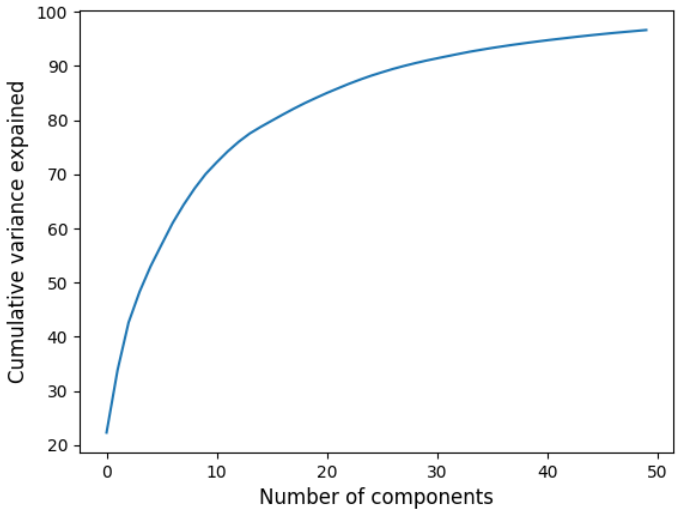

Fig. 2. Principal Component Analysis Results

range of -0.297 to 0.658 (μ). The coefficient score range -/+ 0.3 is considered to contain weak relationship, we used this range to identify the features that are least or not correlated but have an impact on text cohesion. The six (D) text cohesion features *sentence linking*, *order*, *reason* and *purpose*, *opposition*, *temporal* and *quantpro* from TAACO were identified as least correlated features. The number of significant features identified by correlation analysis to represent text cohesion were less as compared to principal components identified by PCA, hence, we used features identified by correlation analysis for further study.

We conducted a survey to identify the impact of these features on answer rating, i.e., on the quality of answer. We shortlisted 3 questions per year (answers without any grammatical errors) along with their answers from 343 sample (S) records. The selected answers were modified on D cohesion features. The modified answer text set was prepared by removing explicit cohesive markers, which makes the answer less accessible as compared to the original answer text. The key intent contained in the answer text was not removed in the iterative process of re-writing the answer text. Apart from removing the explicit connectives and other references, complex sentences were simplified in to multiple sentences with lesser explicit connectives across them. To restrict the role of implicit discourse signals, which activate default choice of continuity [2] between sentences, order of sentences from the original texts were changed. For every iteration of re-writing, the text was validated using TAACO to ensure a difference from the original answer text in terms of cohesion features. While modifying the text, grammar was maintained intact as the focus is evaluating cohesion of the text. The identified cohesion features *sentence linking*, *order*, *opposition* were mapped to the survey question on

reading and *remembering* and cohesion feature *reason* and *purpose* was mapped to the survey question on *understanding*. We did not modify temporal and *quantpro* (quantitative pronouns such as *many*) features in answer as the focus was on *reading*, *remembering* and *understanding*. The 24 original and 24 modified answers (available as Cohesion Survey Questions on Google drive [18]) were shared with Information Security Experts, Computational Linguists and Software Engineers to rate on an ordinal scale (rating from 1 - 6 with value 6 being the highest and 1 being the lowest, we used an even number scale to reduce the chance of participants taking a middle path). Information Security Experts are considered to be contributors/responders in CQA. Software Engineers are considered to be viewers or seekers of responses in CQA. Computational Linguists are expected to be particular about features of text cohesion, are neither consumers or contributors of Information Security CQA. The first 3 questions in the survey capture the participant profile and the next 3 survey response ratings were on correctness, *reading*, *remembering* and *understanding* of the answer. Each participant was provided with 3 StackExchange questions followed by its answer with a mix of original and modified answers. No participant was given the original and modified text of the same answer to avoid bias and to measure responses as independent samples.

- Q1 : Name, Email ID and Educational Qualification of the participant were requested. The purpose of capturing this information was to bring seriousness to the survey and validity to the data for any future reference. The response to this question will not be used for analysis.
- Q2 : How regularly do you use crowdsourcing platforms like StackExchange, Quora and MTurk, etc? This question validates the level of participation on CQA platforms. This question was expected to allude survey participants that authors were interested on CQA quality assessment. However, participants were not informed on the type (cohesion in the text) of quality assessment.
- Q3 : What is your professional/personal familiarity in Information Security domain? The selected StackExchange sample CQA were related to information security, the domain familiarity provides an insight into participants response on correctness of the StackExchange answer.
- R1 : Is the given answer responding to the question? As StackExchange is moderated and we selected answers that have median score 12, answers were expected to be near correct. This survey question was posed to eliminate non-serious survey participants and understand technical challenges of non-security domain participants.
- R2 : What is the ease of reading and remembering the answer? The response to this survey question identify the ease of *reading* and *remembering*, i.e. related to *sentence linking*, *order*, *opposition* cohesion features. We combined the question on reading and remembering together as text is just not a combination of words but words that form a sentence for remembrance.

– R3 : What is the difficulty level of the answer text? The response to this survey question identifies the *understandability*, ability to gather the intent of the text and relates to *reason* and *purpose* cohesion features.

3 Survey Results and Analysis

We used online and physical forms to interact with Group 1 - Subject Matter Experts referred as SMEs (Academicians, Chief Information Security Officers, Research Students with expertise in Information Security), Group 2 - CLs (Computational Linguists from academia) and Group 3 - SEs (industry professionals and research students in Software Engineering) over a period of 6 months (April - September 2018) for the survey. A total of 246 participants provided 697 responses as shown in table 1 with approximately 50% were SMEs, survey responses are available on Google drive[18]. We had minimum 5 responses to each of the original and modified sample answers from each of the groups. In response to $Q2$, 72% of the participants rated between 4 - 6 on their

Group	**Count**	**Response**
Linguists (CL)	76	105
Software Engineers (SE)	36	111
Security Experts (SME)	134	427
Total	**246**	**643**

Table 1. Summary of Survey Participation

participation on crowdsourcing platforms, this confirmed that most of the participants are familiar with StackExchange or other CQA systems. For $Q3$ on familiarity to Information Security domain, figure 3 shows a near normal distribution suggesting that we have SMEs; all CLs and SEs that may not have security knowledge. We observed that participants grouped in SMEs have rated themselves at different levels on security expertise; 13% of the group have rated themselves low on domain expertise. As shown in figure 3, In response to the correctness of the response ($R1$), we observed that about 5% of the responses were rated answer as not correct (rated response between 1 and -3) and we considered these participants as casual and were removed from further analysis, these participants were present in all 3 groups (SE, CL and SME). These participants also gave same rating as $R1$ to $R2$ and $R3$. In response to $R2$, 83% of the responses rated modified text as relatively less readable and difficult to remember, shown in Table 2. Further analysis at individual question also showed that 70% of the participants had difficulty in *reading* and *remembering*. For $R3$, we analyzed the impact of *understandability* with original ($Orig$) and modified (Mod) answers. Approximately 75% of the participants rated original text higher on *understandability* than modified text (relatively less cohesive) as shown in Table 3, cells marked bold in the table

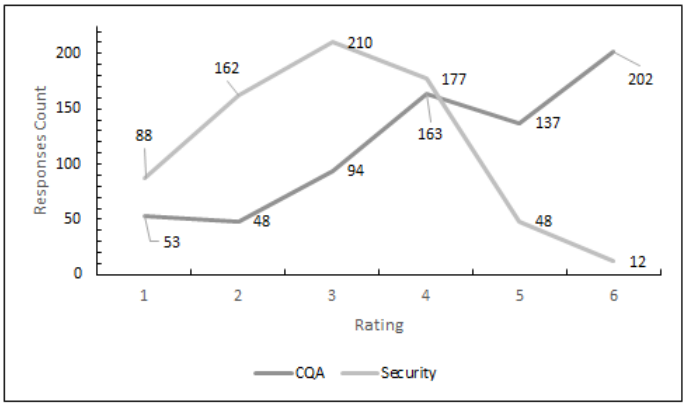

Fig. 3. Participants CQA and Security Familiarity

deviate from the observed pattern. Only 68% of the participants had more than 3 rating and no response had a mean rating of ≥ 5, this states that answers text need to be cohesive for improved understanding. The statistical mode value (variance of $\approx$1.29%) on the rating for all 3 groups is lower for modified answer text as compared to the original answer text. The participants

Rating	**R2**		**R1 on Answer Correctness**					
	Orig	**Mod**	**4**		**5**		**6**	
			Orig	Mod	Orig	Mod	Orig	Mod
1	**23**	**29**	6	4	**7**	**10**	**8**	**11**
2	87	59	20	12	35	23	22	7
3	81	67	**29**	**35**	28	14	8	4
4	91	62	25	16	29	16	15	9
5	43	42	16	14	**9**	**12**	7	3
6	13	5	4	2	2	1	5	1

Table 2. Reading and Remembering of Text

gave a higher mean rating (4) for $R2$ as compared to $R3$ that had a mean of 3 for modified answers. However, rating by SE and SMEs had same mode value for original and modified text. This states that cohesion of text was relatively more important for understanding of the text as compared to reading and remembering. The figure 4 confirms that participants had varying difficulty in *reading*, *remembering* and *understanding* of text though the correctness of the answer text was maintained. We performed ANOVA analysis on the original and modified text in response to $R2$ and $R3$, data is shown in Table 4. The calculated F_{Value} is greater than $F_{Critical}$ score for confidence level of 95%, this confirmed our hypothesis that text cohesion has an impact on rating.

Rating	R3		R1 on Answer Correctness					
	Orig	Mod	4		5		6	
			Orig	Mod	Orig	Mod	Orig	Mod
1	**9**	**12**	2	2	**2**	**5**	2	2
2	53	25	16	4	10	5	8	3
3	73	57	26	20	25	10	3	3
4	96	67	37	27	26	23	17	4
5	75	49	**15**	**16**	36	16	17	5
6	**35**	**51**	**3**	**13**	**12**	**16**	18	18

Table 3. Understandability of Original and Modified Text

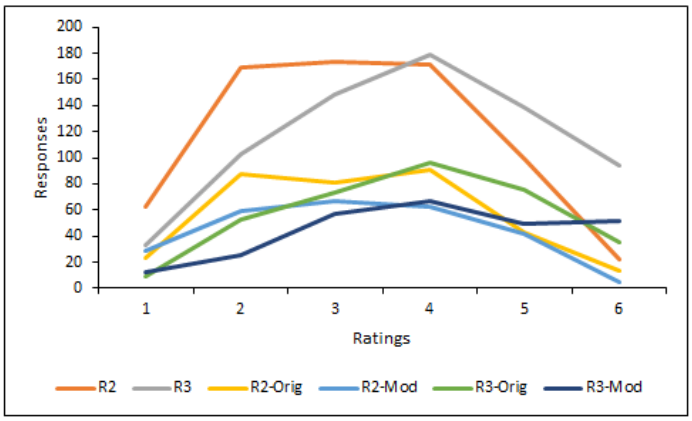

Fig. 4. Distribution of Response Ratings

The $F_{Critical}$ was greater than F_{Value} for $R3$ for SME, however, the P value was 0.17 and greater than α value. This observation reiterated that the text cohesion had an impact on rating for SMEs as well. Based on the ratings, we also observed that text cohesion is more significant for $R2$ (reading and remembering) to SEs and CLs and $R3$ (understanding) to SMEs. The margin of difference of rating between original and modified text is higher for $R3$ as compared to $R2$, similar observations [18] were obtained with one tail and two tail T Test (inferential statistics). Amongst the primary users (SME and SE) of the StackExchange, SMEs gave importance to *reason* and *purpose* to state that responses should be understandable. SEs stated their need of text cohesion is more for *reading* and *remembering* as compared to *understanding*, hence, lack of text cohesion impacts CQA rating. Similar to our experiment, the role of content and domain knowledge was conducted [22] [10] on school children, it showed that background knowledge played a vital role than the reader's decoding skills, and that text cohesion and genre depends on prior knowledge. Their results also revealed that cohesion cues without elaboration information do not facilitate comprehension, particularly for challenging texts. Another

Group	R2		R3	
	F Value	**F Critical**	**F Value**	**F Critical**
SME	0.87	0.78	**1.04**	**1.29**
SE	0.97	0.61	0.78	0.61
CL	**0.66**	**0.68**	0.91	0.67

Table 4. F Scores of R2 and R3

related study [7] described that elaboration and improved cohesion led to higher rating of coherence compared to original and elaborated versions.

4 Conclusion

The empirical study confirmed that cohesion improves *reading*, *remembering* and *understanding* of CQA text and impacts on quality or rating of CQA. Based on the ratings across groups, it is identified that importance of text cohesion is more significant for *understanding* as compared to *reading* and *remembering*. SMEs stated the need of cohesion for understanding whereas SEs stated the need for reading and remembering of the CQA text. Based on these analyses, a plugin for cohesion assessment of the text can be built, thereby, to validate the quality of web text on a page. Based on the ratings from SMEs on importance of cohesion for understandability, we state that validating text cohesion will enhance algorithms that are used for building/enhancing ontologies. The text cohesion may also improve credibility of a web page and assist in identifying fake content. Our experiment was limited to 643 responses obtained from 246 users on a sample 24 questions, extending the sample size or user base may provide more insights on text cohesion. Having traditional linguists as compared to computational linguists as survey participants would strengthen our observations on the need of text cohesion. The survey questions can be extended to open domain to validate if there are any different observations.

References

1. Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H.R., Bertino, E., Dustdar, S.: Quality Control in Crowdsourcing Systems: Issues and Directions. IEEE Internet Computing **17**(2), 76–81 (2013)
2. Ariel, M.: Accessibility Theory: An Overview. Text Representation: Linguistic and Psycholinguistic Aspects **8**, 29–87 (2001)
3. Boehm, B.W., Brown, J.R., Lipow, M.: Quantitative Evaluation of Software Quality. In: Proceedings of the 2nd International Conference on Software Engineering. pp. 592–605. IEEE Computer Society Press (1976)
4. Britton, B.K., Gülgöz, S.: Using Kintsch's Computational Model to Improve Instructional Text: Effects of Repairing Inference Calls on Recall and Cognitive Structures. Journal of Educational Psychology **83**(3), 329 (1991)

5. Burghardt, K., Alsina, E.F., Girvan, M., Rand, W., Lerman, K.: The Myopia of Crowds: Cognitive Load and Collective Evaluation of Answers on Stack Exchange. PloS one **12**(3), e0173610 (2017)
6. Charolles, M., Ehrlich, M.F.: Aspects of Textual Continuity Linguistic Approaches. In: Advances in Psychology, vol. 79, pp. 251–267. Elsevier (1991)
7. Crossley, Scott A.; McNamara, D.S.: Say More and Be More Coherent: How Text Elaboration and Cohesion can Increase Writing Quality. Institute of Educational Sciences (2016)
8. Crossley, S.A., Kyle, K., McNamara, D.S.: The Tool for the Automatic Analysis of Text Cohesion (TAACO): Automatic Assessment of Local, Global, and Text Cohesion. Behavior research methods **48**(4), 1227–1237 (2016)
9. Dahiya, Y., Talukdar, P.: Discovering Response-eliciting Factors in Social Question Answering: A Reddit Inspired Study. Director **24196**(3295), 13–61 (2016)
10. Danielle S McNamara, Ozuru Yasuhiro, R.G.: Comprehension Challenges in the Fourth Grade: The Roles of Text Cohesion, Text Genre, and Readers' Prior Knowledge. International Electronic Journal of Elementary Education (2011)
11. Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z.: Coh-Metrix: Analysis of Text on Cohesion and Language. Behavior Research Methods, Instruments, & Computers **36**(2), 193–202 (2004)
12. Halliday, M.A.K.: Explorations in the Functions of Language. Canadian Journal of Linguistics/Revue canadienne de linguistique **21**(2), 196–199 (1976)
13. Kintsch, W., Van Dijk, T.A.: Toward a Model of Text Comprehension and Production. Psychological Review **85**(5), 363 (1978)
14. Li, B., Jin, T., Lyu, M.R., King, I., Mak, B.: Analyzing and Predicting Question Quality in Community Question Answering Services. In: 21st International Conference on World Wide Web. pp. 775–782. ACM (2012)
15. Liu, J., Shen, H., Yu, L.: Question Quality Analysis and Prediction in Community Question Answering Services with Coupled Mutual Reinforcement. IEEE Transactions on Services Computing **10**(2), 286–301 (2017)
16. McNamara, D.S., Kintsch, W.: Learning from Texts: Effects of Prior Knowledge and Text Coherence. Discourse Processes **22**(3), 247–288 (1996)
17. Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Yang, B., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., et al.: Never-ending Learning. Communications of the ACM **61**(5), 103–115 (2018)
18. for Now, A.: CQA Text Cohesion Analysis. https://tinyurl.com/CQACohesion/ (2019), [Online; accessed 20-Dec-2019]
19. Ratner, B.: The Correlation Coefficient: Its Values Range Between +/-1, or Do They? Journal of Targeting, Measurement and Analysis for Marketing **17**(2), 139–142 (Jun 2009)
20. Ravi, S., Pang, B., Rastogi, V., Kumar, R.: Great Question! Question Quality in Community Q&A. Eighth International AAAI Conference on Weblogs and Social Media **14**, 426–435 (2014)
21. Van Der Maaten, L., Postma, E., Van den Herik, J.: Dimensionality Reduction: A Comparative Review. Journal of Machine Learning Research **10**, 66–71 (2009)
22. William H. Rupley, V.L.W.: Content, Domain, and Word Knowledge : Relationship to Comprehension of Narrative and Expository Text. Reading and Writing (1996)
23. Zowghi, D., Gervasi, V.: The Three Cs of Requirements: Consistency, Completeness, and Correctness. In: International Workshop on Requirements Engineering: Foundations for Software Quality, Essen, Germany: Essener Informatik Beitiage. pp. 155–164 (2002)