What Makes a Good Answer in Community Question Answering? An Analysis of Assessors' Criteria

Daisuke Ishikawa† Noriko Kando† Tetsuya Sakai*
†National Institute of Informatics *Microsoft Research Asia
{dais,kando}@nii.ac.jp, tetsuyasakai@acm.org

ABSTRACT

Community question answering (CQA) has recently become a popular means of satisfying personal information needs, and methods for effectively retrieving information from CQA archives are attracting research interest as the number of reusable questions and answers are rapidly increasing. Rather than having to post a question and wait for an answer, an individual can often obtain an immediate answer by searching the archives. However, as the quality of archived answers varies widely, a method is needed for effectively extracting high-quality answers. In this study, we manually selected random questions and answers from the archives for Yahoo! Chiebukuro, a Japanese CQA site equivalent to Yahoo! Answers, had them evaluated by four assessors, and identified the criteria used by the assessors in their evaluations. These criteria should be useful in constructing a model for identifying high-quality answers.

Keywords: Community QA, Yahoo! Chiebukuro, High-Quality Answer, Kappa Coefficient, Qualitative Data Analysis

1. INTRODUCTION

The construction of a test collection requires human assessment and annotation, and the method and quality of the assessment affect the quality of the test collection. The interpretation and application of the assessment criteria by an assessor is not well understood. Although studies on agreement among assessors and on majority voting evaluation have been reported, there have been no reports on the reasons for disagreement among assessors.

In community question answering (CQA) studies, the criteria for answer quality must be considered as well as the topical relevance. In this study, We address the problem of answer quality in the NTCIR-8 CQA task by focusing on the criteria actually used by each assessor and on the characteristics of a high-quality answer. We investigated the correspondence between the results of the analysis and system evaluation.

We found that the criteria actually used differed among assessors and that one assessor in our study who had used various criteria had the highest performance in system evaluation. We also identified assessment criteria that had not been mentioned in previous reports. These findings should be useful in developing an assessment methodology for constructing a test collection as well as for developing a model of information access systems that can be used to clarify the complex information needs of which users often choose CQA as a venue for searching for relevant information and answers.

The structure of this paper is as follows. The next section describes related work. Section 3 describes the NTCIR-8 CQA test collection used and the assessment procedure. Section 4 describes the agreement among the assessors. Section 5 describes the procedure used for identifying the criteria actually used, the criteria identified, and the effects of these criteria on system effectiveness. Section 6 summarizes the key points and mentions future work.

2. RELATED WORK

Although several methods for automatically identifying high-quality (or best) answers have been reported[1][2][3], there have been no reports on the criteria the assessors actually used. For effective information acquisition from CQA archives, it is important to identify high-quality answers that can be used by future users with information needs similar to those of previous users who originally posted the questions.

Shah and Pomerantz investigated the answer quality in CQA[4] by using five assessors and thirteen criteria[5]: *informative, polite, complete, readable, relevant, brief, convincing, detailed, original, objective, novel, helpful,* and *expert*. In Shah and Pomerantz, these detailed criteria for high-quality answers were given to the assessors by the researchers, so they may not be exhaustive from the assessor's viewpoint.

In this study, we analyzed unstructured natural language reports prepared by four assessors on their assessment of answers taken from a CQA archive and identified the criteria actually used by each assessor. That is, we identified the criteria for high-quality answers from the assessor's viewpoint rather than the researcher's viewpoint.

3. ASSESSMENT OF ANSWER QUALITY

3.1 NTCIR-8 CQA Test Collection

The NTCIR-8 CQA test collection [6] [7] used in the assessment contains 1,500 questions and 7,443 answers associated to these questions, as shown in Table1. These questions were chosen at random from the Yahoo! Chiebukuro data (version 1.0) [8] for the 15 top categories[9] in terms of the number of questions (Table 2).

Table 1: NTCIR-8 CQA test collection [6]

Questions	1500
Answers	7443
(Best answers)	(1500)
(Other answers)	(5943)
Average no. of answers per question	4.962
Data size	4.0 MB (UTF-8 encoding)

"Yahoo! Chiebukuro" data (version 1.0) has been available for

Table 2: Number of questions by category in NTCIR-8 CQA collection

Category	Category (in Japanese)	Number of questions
yahoo	Yahoo! JAPAN	222
entertainment	エンターテインメントと趣味	167
health	健康、美容とファッション	154
lifeguide	暮らしと生活ガイド	136
internet	インターネット、PC と家電	135
sports	スポーツ、アウトドア、車	120
love	生き方と恋愛、人間関係の悩み	120
education	教養と学問、サイエンス	120
school	子育てと学校	89
news	ニュース、政治、国際情勢	63
travel	地域、旅行、お出かけ	58
business	ビジネス、経済とお金	42
career	職業とキャリア	38
manners	マナー、冠婚葬祭	36
Total		1500

research purposes since 2007. Table 3 gives an overview of the data.

Table 3: Overview of "Yahoo! Chiebukuro" data [8]

Date range	2004/4/1 − 2005/10/31
Questions resolved	3,116,009 items (about 916 MB)
Best answers	3,116,008 items (about 935 MB)
Other answers	10,361,777 items (about 2.3 GB)

Both data sets are available for research purpose from the Informatics Data Repository of the National Institute of Informatics [1], Japan.

3.2 Assessment

3.2.1 Assessors' Backgrounds

We hired four university students who had used Yahoo! Chiebukuro to assess the 7443 answers to the 1500 questions. Table 4 summarizes their backgrounds.

3.2.2 CQA Assessment System

The system used for assessment was based on SEPIA (Standard Evaluation Package for Information Access) [10], which was originally developed to use in the data creation, annotation and assessment in the advanced cross-lingual information access (ACLIA) task at NTCIR-7 and -8, and is an open source tool to develop the test collections for information retrieval and question asnwering systems and conduct relevance judgment. The system is described in detail in a previous paper [7].

3.2.3 Assessment Guidelines

The four assessors individually assessed the 7443 answers and graded each one A, B, or C.

- A: Satisfactory answer to the question.
- B: Partially relevant answer to the question.
- C: Unrelated to the question.

[1]NII/IDR, http://http://www.nii.ac.jp/cscenter/idr/en/index.html

The purpose of this study is to clarify the implicit criteria used by users when judging whether the answer is good. To this end, we deliberately left the judgment guideline vague, as shown above.

The answer order for each question was shuffled in the test collection. Assessors could, where appropriate, assign the same grade to every answer for a given question.

3.2.4 Assessment results

Table 5 shows the number of answers by grade and the total assessment time for each assessor.

4. AGREEMENT AMONG ASSESSORS

We calculated the κ coefficient among assessors by using the results from Table 5 [12]:

$$\kappa = \frac{Po - Pc}{1 - Pc}$$

(Po : observed agreements, Pc : chance agreements)

Landis and Koch [13] shows the judgment standards for the κ coefficient as follows:

- $0.0 \sim 0.2$: slight
- $0.21 \sim 0.4$: fair
- $0.41 \sim 0.6$: moderate
- $0.61 \sim 0.8$: substantial
- $0.81 \sim 1.0$: almost perfect

Table 6 summarizes the κ coefficients between assessors and by category.

From these results, we see that

- The κ coefficients by category for Assr-2 $\leftrightarrow$ 3 and Assr-1 $\leftrightarrow$ 4 were generally higher than those for the other assessor pairs.
- The difference in the κ coefficients was due to the differences among the categories (for example, approximately 0.30-0.45, or fair/moderate agreement, for the category "education" whereas 0.03-0.33, or slight/fair agreement, for the category "love").

Table 4: Assessors' backgrounds

	Assr-1	Assr-2	Assr-3	Assr-4
Gender	male	male	female	female
Age	20s	20s	20s	30s
Undergraduate year	3rd	4th	3rd	3rd
Field	Sciences	Arts	Sciences	Arts
Major field of study	Informatics (Simulation)	Literary theory	Biochemistry	Asian history
Frequency of using Yahoo! Chiebukuro	once in two weeks	twice or three times a week	twice or three times a month	once a month
Posted question/answer?	no	no	no	no

Table 5: Assessment results

Assessor	Grade A	Grade B	Grade C	Total time (hours)
Assr-1	4879	2420	144	15.96
Assr-2	2311	4967	165	23.54
Assr-3	2757	4399	287	25.11
Assr-4	4327	2996	120	21.20

5. IDENTIFICATION OF CRITERIA

To identify the criteria actually used by the assessors, we analyzed their reports using qualitative data analysis.

Computer-assisted qualitative data analysis software (CAQDAS) such as Atlas.ti, MaxQDA, NVivo, and Weft QDA supports the manaual qualitative data analysis by human analysts[14]. We used Weft QDA [15], which is open source, because it is easy to use and sufficiently robust for our needs.

The setting for the analysis was as follows

- CAQDAS: Weft QDA version 1.0.1 (for Windows)
- Target sources: text files
- Data size: Assr-1 (6716 bytes), Assr-2 (3100 bytes), Assr-3 (4925 bytes), Assr-4 (6225 bytes)

The procedure was as follows.

1. Import assessors' reports (text files).
2. Read assessors' reports and find descriptions related to criteria.
3. Create new criteria name (temporary name) on basis of description.
4. Use criteria (temporary) to mark descriptions related to criteria.
5. Repeat steps 2-4 until new criteria are not found.
6. Review descriptions marked as criteria and change criteria name to suitable name.
7. Count number of criteria used by each assessor.

5.1 Identified criteria

Using the procedure above, we identified 12 criteria that the assessors used to assess the quality of the answers.

1. **[experience]** Does the answerer discuss his own experience?
2. **[evidence]** Does the answer contain supporting evidence?
3. **[politeness]** Is the answer polite?
4. **[detail]** Is the answer detailed?
5. **[opinion]** Does the answer reflect a personal opinion?
6. **[relevance]** Does the answer actually answer the question?
7. **[concreteness]** Does the answer contain concrete examples or instructions?
8. **[respect]** Does the answerer respect the questioner's feelings or situation?
9. **[logic]** Is the answer logical?
10. **[reason]** Does the answer explain why it is correct?
11. **[trustworthiness]** Does the answer appear trustworthy?
12. **[persuasiveness]** Is the answer persuasive?

The total results of criteria used by each assessor are shown in table7.

Several observations can be made from these results:

- Assr-3 assessed the quality of the answers on the basis of a wide variety of criteria.
- Assr-2 assessed the quality of the answers on the basis of a narrow variety of criteria (mainly "experience", "evidence", and "politeness").
- Assr-1 mainly used "opinion" and Assr-4 mainly used "relevance" for their assessments. However, these criteria are weaker than the other criteria.
- The κ coefficients between Assr-1 and Assr-4 and between Assr-2 and Assr-3 are higher than the others. However, the combination of criteria that each these assessors used was different.
- The "concreteness", "respect", "logic", and "trustworthiness" criteria that Assr-3 used are rarely mentioned in previous reports such as [4] and [5].
- The "logic" and "trustworthiness" criteria were selectively used in situations in which other criteria could not be used.

In short, each assessor assessed the quality of the answers from considerably different viewpoints.

Table 6: Kappa coefficients between assessors

(common ground)	Assr-1 ↔ 2 (male)	Assr-3 ↔ 4 (female)	Assr-1 ↔ 3 (Sciences)	Assr-2 ↔ 4 (Arts)	Assr-1 ↔ 4 (none)	Assr-2 ↔ 3 (none)
No. of answers graded the same A	2072	2255	2473	1962	3574	1531
No. of answers graded the same B	2082	2286	1938	2543	1616	3535
No. of answers graded the same C	55	51	39	50	42	77
Po	0.565	0.616	0.597	0.611	0.702	0.690
Pc	0.420	0.453	0.435	0.449	0.512	0.510
κ	0.249	0.298	0.287	0.295	0.390	0.368
κ coefficient by category						
yahoo	0.247	0.214	0.266	0.213	0.306	0.378
entertainment	0.385	0.167	0.238	0.354	0.374	0.392
health	0.15	0.222	0.091	0.208	0.356	*0.441
lifeguide	0.244	0.317	0.255	0.395	0.408	0.37
internet	0.278	0.35	0.391	0.347	0.402	0.335
sports	0.264	0.376	0.254	0.391	0.401	*0.465
love	0.028	0.326	0.137	0.185	0.198	0.164
education	*0.453	0.313	0.4	0.298	*0.412	*0.471
school	0.159	0.337	0.331	0.21	0.361	0.247
news	0.244	0.286	0.287	0.256	0.346	0.311
travel	0.369	0.175	0.286	0.387	0.315	0.372
business	0.369	0.247	0.209	0.336	0.402	0.311
career	0.208	0.263	0.161	0.325	*0.555	*0.431
manners	0.138	0.118	0.089	0.144	*0.429	*0.433

κ coefficients > 0.41 are marked by $*$

5.2 Effects on System Effectiveness

Based on their assessments, we constructed GAW (Good Answers with Weights) data that assigned a judgment weight to each label A, B and C. Details of GAW data were described in our paper [11]. Using the GAW data, we propose to compute three graded-relevance evaluation metrics: *GAW-nG@1*, *GAW-nDCG* and *GAW-Q*. In our study, we evaluated these assessor assessments. Table8 summarizes our findings.

The "COM" entries in the table 8 show the results for answer quality evaluated by a computer. The "BASELINE" entries show the baseline evaluation results: "BASELINE1" is for answers ranked randomly, "BASELINE2" is for answers ranked by length (the longer is the better), and "BASELINE3" is for answers ranked by timestamp (the newer is the better).

"BA-Hit@1" is an evaluation based on the best answer selected by the original asker of the question, and "GAW-nG@1" is an evaluation based on the good quality answers assessed by the assessors. Both evaluations evaluate whether the best answer or most good answer was chosen. The "GAW-nDCG" and "GAW-Q" evaluations are respectively based on evaluation metrics nDCG [16] and Q-measure[17] used in information retrieval. Both evaluate whether the answers are sorted in order of quality.

In general, BA data may be biased and nonexhaustive, as the judgments basically rely on a single person and there is only one BA per question [11]. The GAW data based on the four assessors were constructed for alleviating these issues.

GAW-nG@1 evaluates exactly one answer returned by the system (or the assessor), i.e. one that the system considers to be the best among the posted answers. Whereas, GAW-nDCG and GAW-Q evaluate a ranked list of answers, i.e. the ability of the system (or the assessor) to determine the relative quality of answers.

Table 8 (duplicated from [11]) shows that assessor performance was in the order

$$Assr-3 > Assr-2 > Assr-1 > Assr-4.$$

There was a great difference, in particular, between the performances of Assr-2 and Assr-1. Assr-3 not only outperformed the other assessors, she also outperformed the computer evaluations, except the BA-Hit@1 evaluation.

Among the criteria identified in this study, some may have a stronger impact than others. For example, "relevance" and "opinion" are weak criteria in this study because many answers actually answer the question, and have written own opinion. On the other hand, "evidence" and "concreteness" are thought to be strong criteria because few answers include evidences, and are described concretely.

Assr-3 uses strong criteria and weak criteria properly according to the situation. Assr-2 chiefly uses strong criteria. Assr-1 and 4 chiefly use weak criteria. These different use of the criteria may explain the asessors' performance differences shown in Table 8.

In short, the criteria and method used by Assr-3 resulted in the most effective assessment of the answers. However, their performances may have been overestimated since the GAW data set was constructed based on judgments from these very judges. To cancel out this effect, we have also conducted leave-one-out experiments for evaluating the assessors and the systems elsewhere [11].

6. CONCLUSION

Our analysis of the criteria used by four assessors to assess the quality of community-answered questions identified twelve criteria used by the assessors. The criteria and method used by one of the assessors were especially effective compared with other human assessors and various computer evaluations. To our knowledge, four of the twelve criteria, namely, "concreteness", "respect", "logic", and "trustworthiness" have not been mentioned in previous work.

Table 7: Total results of criteria used by each assessor

Criteria	Assr-1	Assr-2	Assr-3	Assr-4	total
experience	3	5	1	3	12
evidence	-	3	1	3	7
politeness	-	5	1	-	6
detail	-	-	4	1	5
opinion	3	-	1	1	5
relevance	-	-	-	4	4
concreteness	-	1	2	-	3
respect	1	-	2	-	3
logic	-	-	2	-	2
reason	-	-	2	-	2
trustworthiness	-	-	1	-	1
persuasiveness	-	-	-	1	1

Table 8: Mean performances based on BA and GAW data [11]

BA-Hit@1		GAW-nG@1		GAW-nDCG		GAW-Q	
COM-M2	0.4980	**Assr-3**	0.9567	**Assr-3**	0.9857	**Assr-3**	0.9760
COM-M1	0.4980	**Assr-2**	0.9446	COM-M2	0.9797	COM-M2	0.9688
COM-M4	0.4847	COM-M2	0.9288	COM-M4	0.9795∗∗	COM-M4	0.9687
BASELINE2	0.4847	COM-M1	0.9278	**Assr-2**	0.9794‡	COM-A2	0.9683∗
COM-A2	0.4840	COM-M4	0.9276	COM-M1	0.9791†	COM-M1	0.9679
COM-M3	0.4813	COM-A2	0.9251	COM-A2	0.9790∗∗	COM-A1	0.9674
COM-A1	0.4813†	BASELINE2	0.9242	COM-A1	0.9785	BASELINE2	0.9673∗∗
Assr-3	0.4353	COM-A1	0.9238∗∗	BASELINE2	0.9784∗∗	**Assr-2**	0.9646
Assr-2	0.4187	COM-M3	0.9076∗∗	COM-M3	0.9744∗∗	COM-M3	0.9609∗∗
BASELINE3	0.3820	**Assr-1**	0.8916	**Assr-1**	0.9724	**Assr-1**	0.9573
Assr-1	0.3280∗	**Assr-4**	0.8814∗	**Assr-4**	0.9699∗∗	**Assr-4**	0.9538∗∗
Assr-4	0.3020	BASELINE3	0.8460∗∗	BASELINE3	0.9576∗∗	BASELINE3	0.9366∗∗
BASELINE1	0.2713∗∗	BASELINE1	0.8057∗∗	BASELINE1	0.9455∗∗	BASELINE1	0.9172
COM-L3	0.1767	COM-L3	0.7354	COM-L2	0.9365∗∗	COM-L2	0.9094∗∗
COM-L2	0.1767	COM-L2	0.7354	COM-L3	0.9325∗∗	COM-L3	0.9015∗∗
COM-L1	0.1767	COM-L1	0.7354	COM-L1	0.9291	COM-L1	0.8944

The runs are sorted by performance metric. "∗" and "∗∗" indicate that a run significantly outperformed the one shown immediately below according to a two-sided sign test ($\alpha = 0.05$ and $\alpha = 0.01$, respectively), whereas "†" and "‡" indicate that a run significantly *underperformed* the one shown below ($\alpha = 0.05$ and $\alpha = 0.01$, respectively), contrary to the mean performance ranking. Note that statistical significance is not transitive.

Based on our findings, we plan to construct a model for automatically estimating anwer quality and evaluate its effectiveness.

7. ACKNOWLEDGMENTS

'Yahoo! Chiebukuro Data' provided to the National Institute of Informatics by Yahoo Japan Corporation were used to construct the test collection.

8. REFERENCES

[1] Agichtein, E., Castillo, C., Donato, D., Gionis, A. and Mishne, G.: Finding high-quality content in social media, Proceedings of the International Conference on Web Search and Web Data Mining, pp. 183–194, 2008.

[2] Wang, X.J., Tu, X., Feng, D. and Zhang, L.: Ranking community answers by modeling question-answer relationships via analogical reasoning, Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 179–186, 2009.

[3] Blooma, M.J., Chua, A.Y.K. and Goh, D.H.L.: Selection of the Best Answer in CQA Services, 2010 Seventh International Conference on Information Technology, pp. 534–539, IEEE, 2010.

[4] Shah, C. and Pomerantz, J.: Evaluating and predicting answer quality in community QA, Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 411–418, ACM, 2010.

[5] Zhu, Z., Bernhard, D. and Gurevych, I.: A Multi-Dimensional Model for Assessing the Quality of Answers in Social Q&A Sites, Proceedings of the 14th International Conference on Information Quality (ICIQ 2009), pp. 264–265, 2009.

[6] NTCIR-8 CQA (Community QA) Research Purpose Use of Test Collection: http://research.nii.ac.jp/ntcir/permission/ntcir-8/perm-ja-CQA.html

[7] Ishikawa, D., Sakai, T. and Kando, N.: Overview of the NTCIR-8 Community QA Pilot Task (Part I): The Test Collection and the Task, Proceedings of NTCIR-8 Workshop Meeting, pp. 421-432, 2010.

[8] Distribution of "Yahoo! Chiebukuro" data: http://research.nii.ac.jp/tdc/chiebukuro.html

[9] Yahoo! JAPAN: Yahoo! Chiebukuro: http://chiebukuro.yahoo.co.jp/

[10] SEPIA (Standard Evaluation Package for Information

Access):
http://sourceforge.net/projects/opensepia/

[11] Sakai, T., Ishikawa, D., Kando, N., Seki, Y., Kuriyama, K. and Lin, C.-Y.: Using Graded-Relevance Metrics for Evaluating Community QA Answer Selection, Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM), pp. 187–196, 2011.

[12] Sim, J. and Wright, C.C., The kappa statistic in reliability studies: use, interpretation, and sample size requirements, Physical Therapy, Vol. 85, No. 3, pp. 257–268, 2005.

[13] Landis, J.R. and Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics, Vol.33, No.1, pp.159-174, 1977.

[14] Sinkovics, R.R., Penz, E. and Ghauri, P.N.: Enhancing the trustworthiness of qualitative research in international business, Management International Review, Springer, Vol. 48, No. 6, pp. 689–714, 2008.

[15] Weft QDA: http://www.pressure.to/qda/

[16] Järvelin, K. and Kekäläinen, J.: Cumulated Gain-Based Evaluation of IR Techniques, *ACM TOIS*, Vol. 20, No. 4, pp. 422–446, 2002.

[17] Sakai, T.: On Penalising Late Arrival of Relevant Documents in Information Retrieval Evaluation with Graded Relevance, *Proceedings of the First Workshop on Evaluating Information Access (EVIA 2007)*, pp. 32–43, 2007.