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

Daisuke Ishikawa†, Noriko Kando†, and Tetsuya Sakai*

†National Institute of Informatics
*Microsoft Research Asia

Outline

- **Introduction**
- Assessment of Answer Quality
- Agreement among assessors
- Identification of Criteria
- Effects on System Effectiveness
- Conclusion

Introduction
- Background -

- Construction of test collection requires human assessment.
- Interpretation of assessment criteria by assessor not well understood.
- No reports on criteria assessors actually used.

Introduction
- Related Work -

- Shah and Pomerantz used five assessors and thirteen criteria:
  - *informative, polite, complete, readable, relevant, brief, convincing, detailed, original, objective, novel, helpful, and expert.*

- They may not be exhaustive from assessor's viewpoint.

Introduction

- Motivation & Study Design -

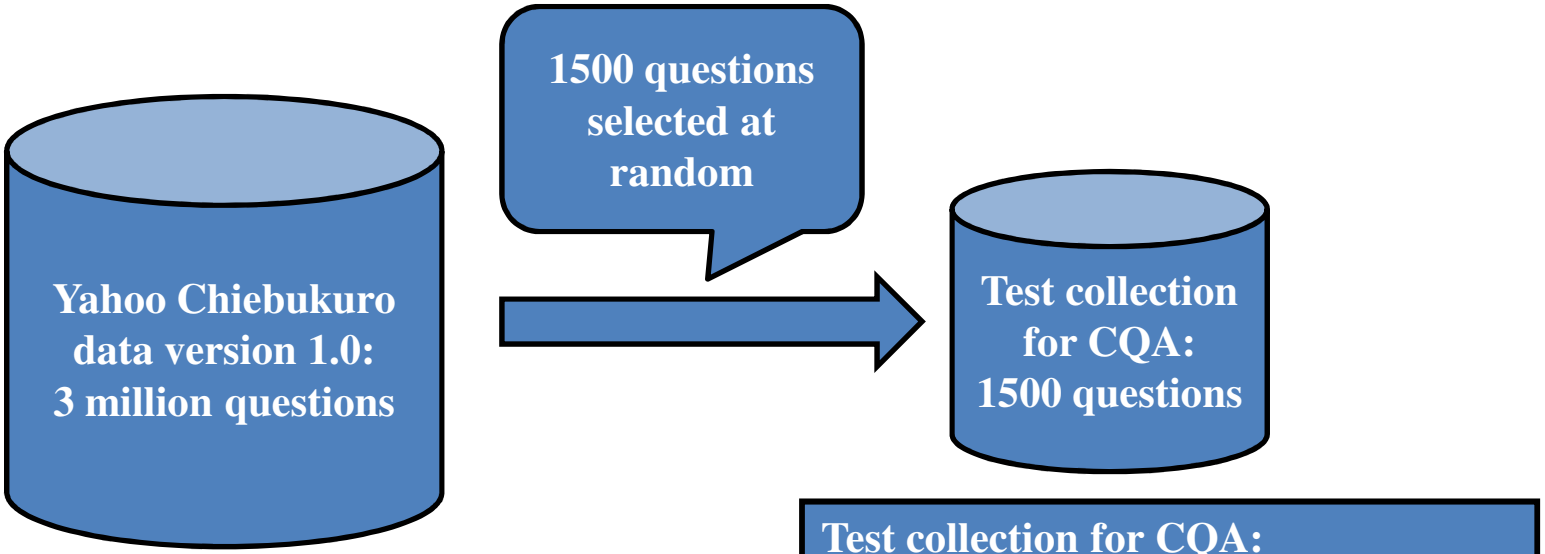
- This study focuses on criteria actually used by each assessor.
- Analyzed unstructured natural language reports written by assessors.
- Using qualitative data analysis.

- Identified criteria from assessor's viewpoint rather than researcher's viewpoint.

Outline

- Introduction
- **Assessment of Answer Quality**
- Agreement among assessors
- Identification of Criteria
- Effects on System Effectiveness
- Conclusion

Assessment of Answer Quality
- NTCIR-8 CQA Test Collection -



Test collection for CQA:

- Questions: 1500
- Answers: 7443
  - ✓ Best answers: 1500
  - ✓ Normal answers: 5943
- Data size: 4.0 MB (UTF-8 encode)

Assessment of Answer Quality
- Assessors' Backgrounds -

Assessor ID	Assr-1	Assr-2	Assr-3	Assr-4
Gender	Male	Male	Female	Female
Age	Twenties	Twenties	Twenties	Thirties
Undergraduate	3rd yr.	4th yr.	3rd yr.	3rd yr.
Arts / Sciences	Sciences	Arts	Sciences	Arts
Major field of study	Informatics (Simulation)	Literary thoughts	Biochemistry	Asian history
Use frequency of Yahoo! Chiebukuro	Once every two weeks	Two or three times per week	Two or three times per month	Once per month
Posting question / answer	Not used	Not used	Not used	Not used

Assessment of Answer Quality
- Assessment Guidelines -

- Assessors graded each one A, B, or C:
  - Grade A: Satisfactory answer to question.
  - Grade B: Partially relevant answer to question.
  - Grade C: Unrelated to question.

Assessment of Answer Quality
- Assessment Results -

Assessor	Grade A	Grade B	Grade C
Assr-1	4879	2420	144
Assr-2	2311	4967	165
Assr-3	2757	4399	287
Assr-4	4327	2996	120

Assessment of Answer Quality
- Assessment Results -

Assessor	Grade A	Grade B	Grade C
Assr-1	4879	2420	144
Assr-2	2311	4967	165
Assr-3	2757	4399	287
Assr-4	4327	2996	120

Assessment of Answer Quality
- Assessment Results -

Assessor	Grade A	Grade B	Grade C
Assr-1	4879	2420	144
Assr-2	2311	4967	165
Assr-3	2757	4399	287
Assr-4	4327	2996	120

Assessment of Answer Quality
- Assessment Results -

Assessor	Grade A	Grade B	Grade C
Assr-1	4879	2420	144
Assr-2	2311	4967	165
Assr-3	2757	4399	287
Assr-4	4327	2996	120

Outline

- Introduction
- Assessment of Answer Quality
- **Agreement among assessors**
- Identification of Criteria
- Effects on System Effectiveness
- Conclusion

Agreement among 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 (moderate) are marked *

Agreement among 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 (moderate) are marked *

Agreement among 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 (moderate) are marked *

Outline

- Introduction
- Assessment of Answer Quality
- Agreement among assessors
- **Identification of Criteria**
- Effects on System Effectiveness
- Conclusion

Identification of Criteria

- Qualitative Data Analysis and Software -

- We analyzed their reports using qualitative data analysis to identify criteria used.
- Computer-assisted qualitative data analysis software (CAQDAS): Atlas.ti, MaxQDA and Weft QDA

- We used Weft QDA:
  - CAQDAS: Weft QDA version 1.01.
  - Target source: text data
  - Data size: Assr-1 (6716 bytes), Assr-2 (3100 bytes), Assr-3(4925 bytes), and Assr-4 (6225 bytes)

Identification of Criteria

- Procedure -

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, 3, 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.

Identification of Criteria

- Identified Criteria -

1. [**experience**] Does answerer discuss his own experience?
2. [**evidence**] Does answer contain supporting evidence?
3. [**politeness**] Is answer polite?
4. [**detail**] Is answer detailed?
5. [**opinion**] Does answer reflect personal opinion?
6. [**relevance**] Does answer actually answer question?

Identification of Criteria

- Identified Criteria -

7. [**concreteness**] Does answer contain concrete examples or instructions?
8. [**respect**] Does answerer respect questioner's feelings or situation?
9. [**logic**] Is answer logical?
10. [**reason**] Does answer explain why it is correct?
11. [**trustworthiness**] Does answer appear trustworthy?
12. [**persuasiveness**] Is answer persuasive?

Identification of Criteria

- Identified Criteria -

7. [**concreteness**] Does answer contain concrete examples or instructions?
8. [**respect**] Does answerer respect questioner's feelings or situation?
9. [**logic**] Is answer logical?
10. [**reason**] Does answer explain why it is correct?
11. [**trustworthiness**] Does answer appear trustworthy?
12. [**persuasiveness**] Is answer persuasive?

Identification of Criteria

- 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

Identification of Criteria

- 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

Identification of Criteria

- 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

Identification of Criteria

- 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

Outline

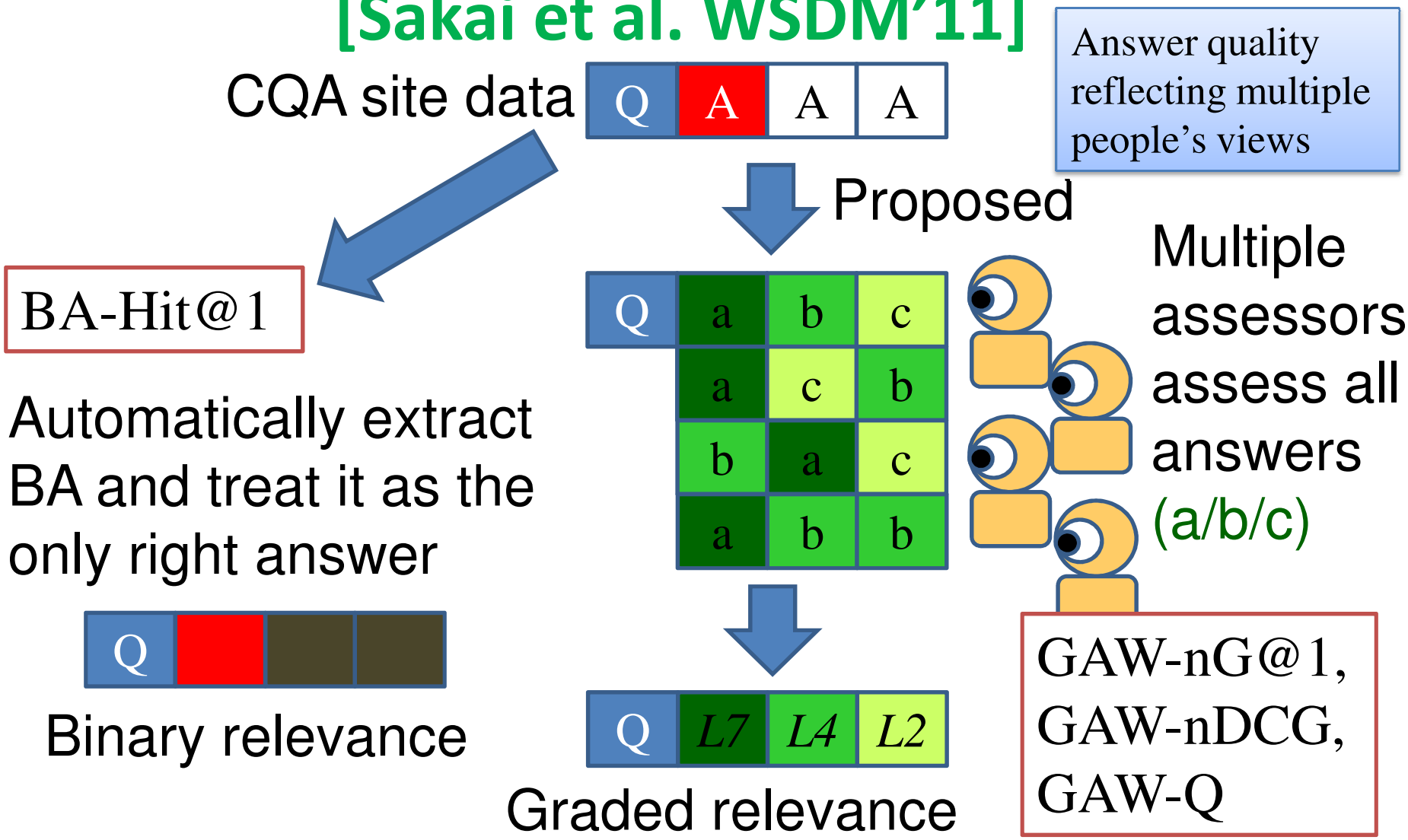
- Introduction
- Assessment of Answer Quality
- Agreement among assessors
- Identification of Criteria
- **Effects on System Effectiveness**
- Conclusion

Effects on System Effectiveness

- We constructed Good Answers with Weights (GAW) data.
- We propose three graded-relevance evaluation metrics:
  - GAW-nG@1
  - GAW-nDCG
  - GAW-Q

Evaluation using multiple assessors and graded relevance metrics

[Sakai et al. WSDM'11]



Effects on System Effectiveness

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

two-sided sign test (* = 0.05 , ** =0.01)

Effects on System Effectiveness

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

two-sided sign test (* = 0.05 , ** =0.01)

Effects on System Effectiveness

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

two-sided sign test (* = 0.05 , ** =0.01)

Assr-3 > Assr-2 > Assr-1 > Assr-4

Effects on System Effectiveness

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

two-sided sign test (* = 0.05 , ** =0.01)

Assr-3 > Assr-2 > Assr-1 > Assr-4

Effects on System Effectiveness

- Why was assessor 3 excellent? -

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

Assr-3 > Assr-2 > Assr-1 > Assr-4

Strong Criteria

Weak Criteria

Outline

- Introduction
- Assessment of Answer Quality
- Agreement among assessors
- Identification of Criteria
- Effects on System Effectiveness
- **Conclusion**

Conclusion

- Identified twelve criteria used by four assessors to assess answer quality of CQA.
- [concreteness], [respect], [logic], and [trustworthiness] criteria are rarely mentioned.
- Assessor 3 was especially effective compared to other assessors and various computer evaluations.

- Based on our findings, we plan to construct a new model for automatically estimating answer quality.

Thank you for your attention.

Daisuke Ishikawa
National Institute of Informatics
dais@nii.ac.jp