Using Graded-Relevance Metrics for Evaluating Community QA Answer Selection

Tetsuya Sakai (MSRA, China)
Daisuke Ishikawa (NII, Japan)
Noriko Kando (NII, Japan)
Yohei Seki (University of Tsukuba, Japan)
Kazuko Kuriyama (Shirayuri College, japan)
Chin-Yew Lin (MSRA, China)

WSDM2011 oral presentation, 11th February , 2011, Hong Kong.

TALK OUTLINE

1. MOTIVATION
2. PROPOSAL
3. EXPERIMENTS
4. CONCLUSIONS

Community QA Data

Q: Is there a good Japanese restaurant in Beijing?

A1: Sobajin in 21st Century Hotel serves the best soba!

A2: Yes.

A3: Miyamotoya in LiangMaQiao serves good yakitori and sake!

A4: There’s no such thing as a good Japanese restaurant!

A5: Takakura’s sushi is great but bloody expensive!

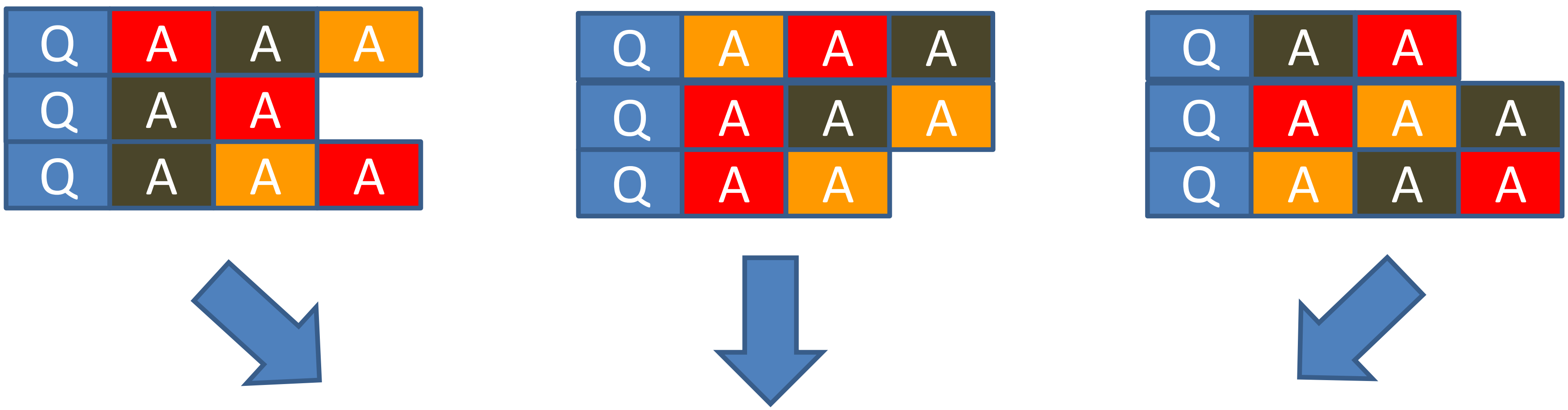
Best Answer (BA)!

Build a system that refines and *reuses* CQA Data

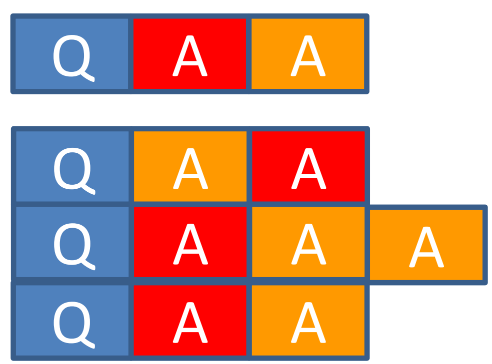
CQA site 1

CQA site 2

CQA site 3



(Consolidate different formats and similar questions)



Select good answers, not just the askers' "best" answers (BAs)!

Problems with previous evaluation of good answer selection from CQA

Q: Is there a good Japanese restaurant in Beijing?

A5: Takakura's sushi is great but bloody expensive!

Only BAs are used for training and evaluation. But BAs may be *biased* (other people might disagree) and/or
nonexhaustive (other answers might also be good)!

A1: Sobajin in 21st Century Hotel serves the best soba!

A3: Miyamotoya in LiangMaQiao serves good yakitori and sake!

Previous Work (selected)

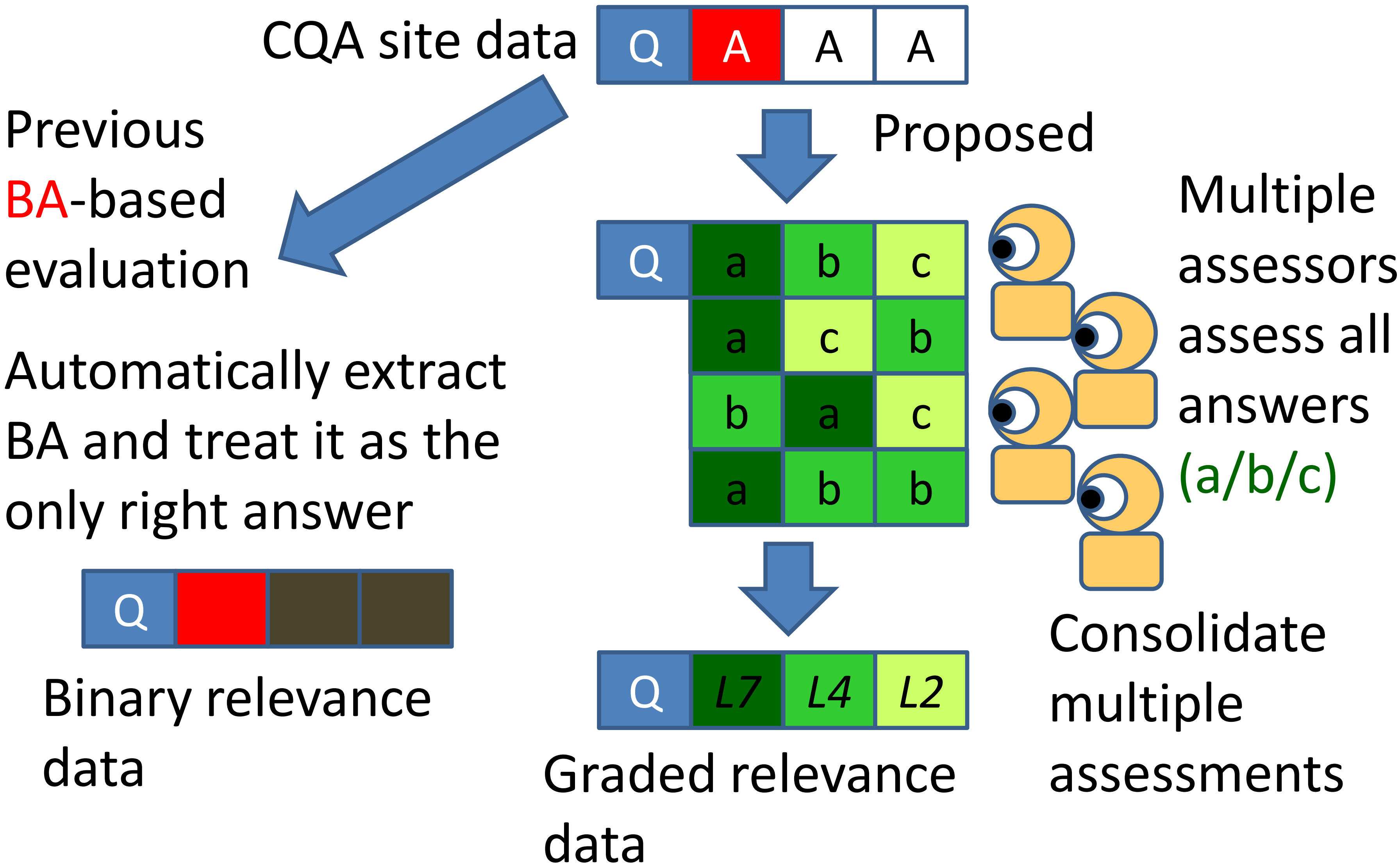
- Precision, Recall, F1 for asker satisfaction prediction [Agichtein/Liu09]
- Average Precision based on three separate gold standards for Q-A ranking [Jeon et al.06]
- Precision based on good quality, relevance and combined [Suryanto et al.09]
- Precision and Reciprocal Rank based on BAs for answer ranking [Wang et al.09]

Everybody uses binary relevance metrics

TALK OUTLINE

1. MOTIVATION
2. PROPOSAL
3. EXPERIMENTS
4. CONCLUSIONS

Hire multiple assessors



Yahoo! Chiebukuro / NTCIR-8 CQA data

- 1,500 *resolved* questions
- 7,443 answers including
  1,500 BAs
- #answers/Q:
  4.96 on average,
  [2,19] range
- 14 Q categories

Grades by 4 judges

Mapped to 9-point relevance

(a) pattern	(b) #answers	(c) weight	(d) level
AAAA	1301	8	$L8$
AAAB	1505	7	$L7$
AABB	1525	6	$L6$
AAA	2	6	
ABBB	1385	5	$L5$
AAB	14	5	
BBBB	1241	4	$L4$
ABB	76	4	
AA	1	4	
BBB	231	3	$L3$
AB	7	3	
BB	105	2	$L2$
A	1	2	
B	32	1	$L1$
(C's only)	17	0	$L0$
total	7443		total

Use graded relevance metrics

- Task: Find one most highly relevant A to a Q

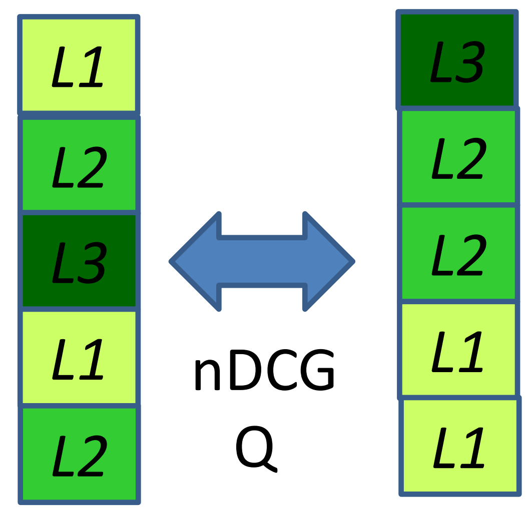
- Normalised Gain at 1 (nG@1)

- Task: Rank all As in order of relevance to Q

- Normalised Discounted Cumulative Gain (nDCG) [Jarvelin/Kekalainen00]
- Q-measure [Sakai04]

System's ranked list of answers

Ideal ranked list of answers



nDCG vs Q-measure

$$nDCG = \frac{\sum_{r=1}^{l} g(r)/\log(r+1)}{\sum_{r=1}^{l} g^*(r)/\log(r+1)} \quad Q = \frac{1}{R}\sum_{r} I(r)\frac{C(r)+\beta cg(r)}{r+\beta cg^*(r)}$$

- nDCG is the *de facto* standard.
- But Q is also a reliable graded relevance metric.
- Highly correlated with nDCG [Sakai IPM07]
- At least as discriminative as nDCG [Sakai SIGIR06]
- Has a user model [Sakai and Robertson EVIA08]
- Reduces to Average Precision [Sakai AIRS04]

TALK OUTLINE

1. MOTIVATION
2. PROPOSAL
3. EXPERIMENTS
4. CONCLUSIONS

Purpose of the NTCIR-8 experiments

Compare BA-based evaluation with proposed methods using multiple assessors and graded relevance

BA-based (binary relevance, only one correct):

- BA-Hit@1 (1 if top-ranked answer is the BA)

Proposed (graded relevance, multiple correct):

- nG@1
- nDCG, Q-measure [assess entire ranked lists]

Ranking runs/Ranking questions

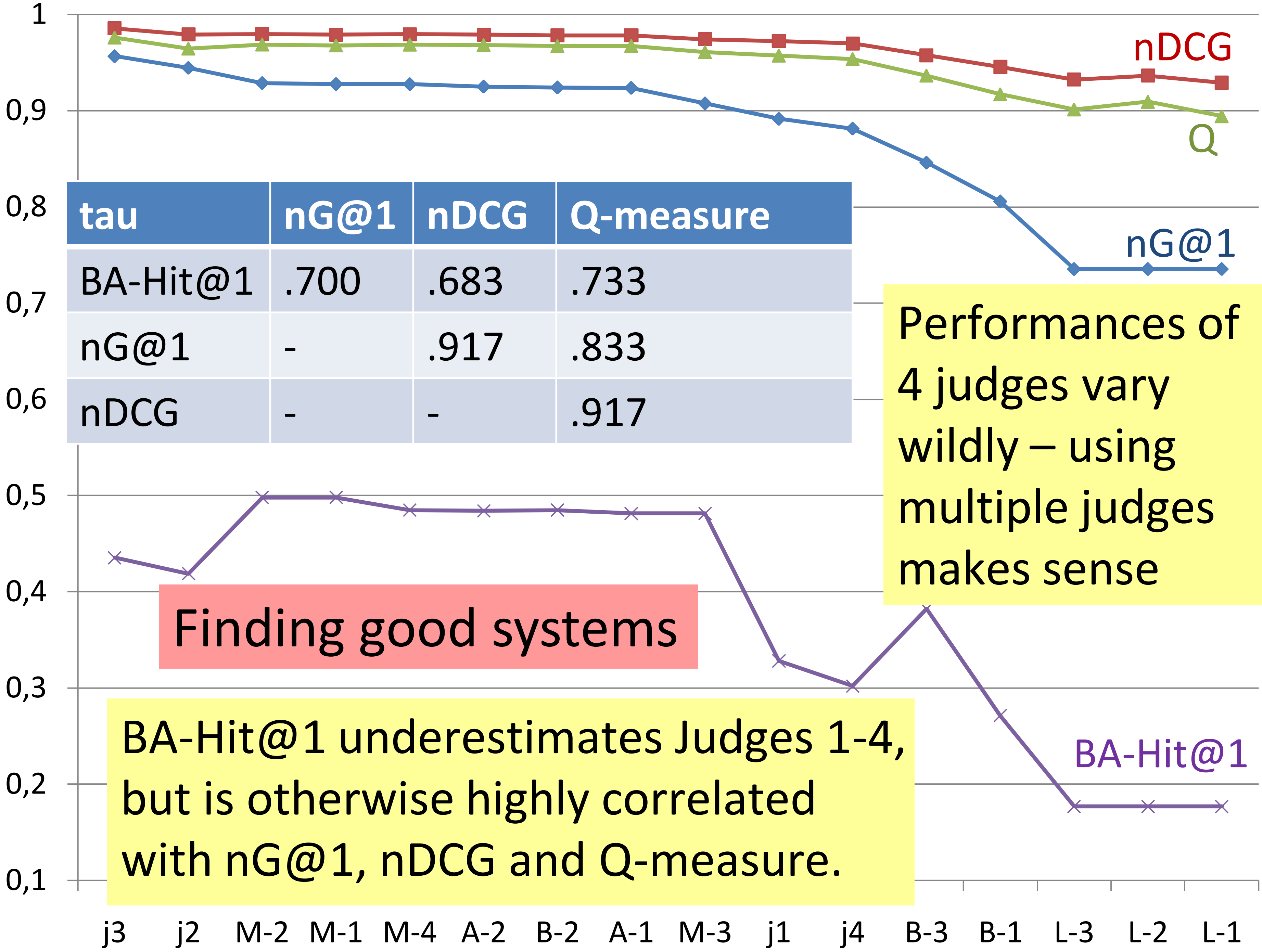
- Find good systems: Ranking 12 runs, plus 4 judges treated as answer rankers (sort answers by a/b/c) by

- *Mean* BA-Hit@1, nG@1, nDCG and Q-measure

over 1500 questions

- Find hard questions: Ranking 1500 Qs by

- *Average* BA-Hit@1, nG@1, nDCG and Q-measure over 12 runs



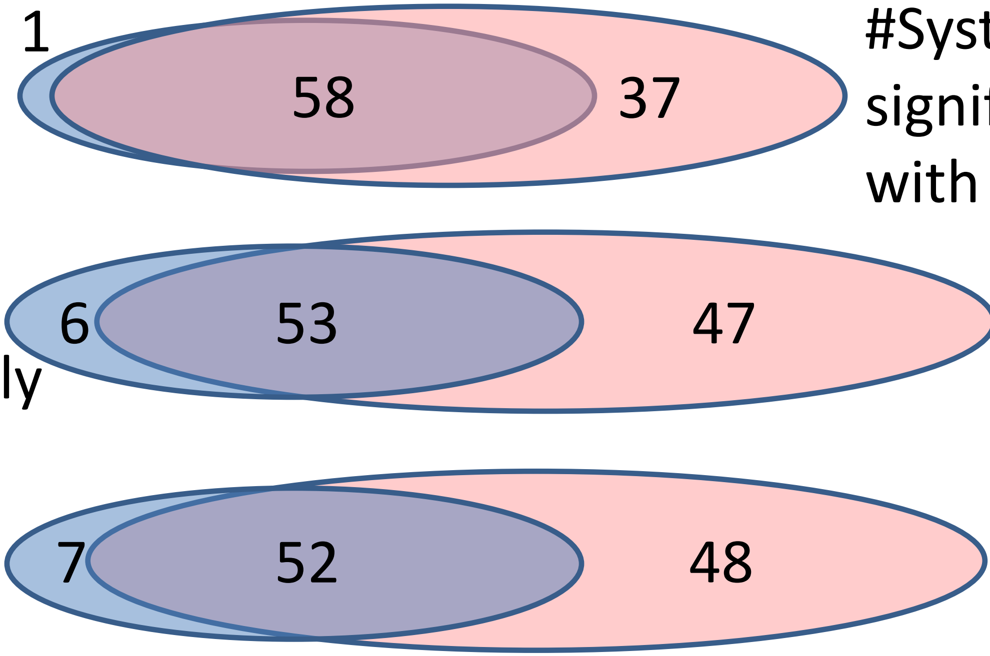
tau	nG@1	nDCG	Q-measure
BA-Hit@1	.700	.683	.733
nG@1	-	.917	.833
nDCG	-	-	.917

Performances of 4 judges vary wildly – using multiple judges makes sense

Finding good systems

BA-Hit@1 underestimates Judges 1-4, but is otherwise highly correlated with nG@1, nDCG and Q-measure.

Discriminative power [Sakai SIGIR06]



#System pairs significantly different with nG@1=95

#System pairs significantly different with BA-Hit@1 =59

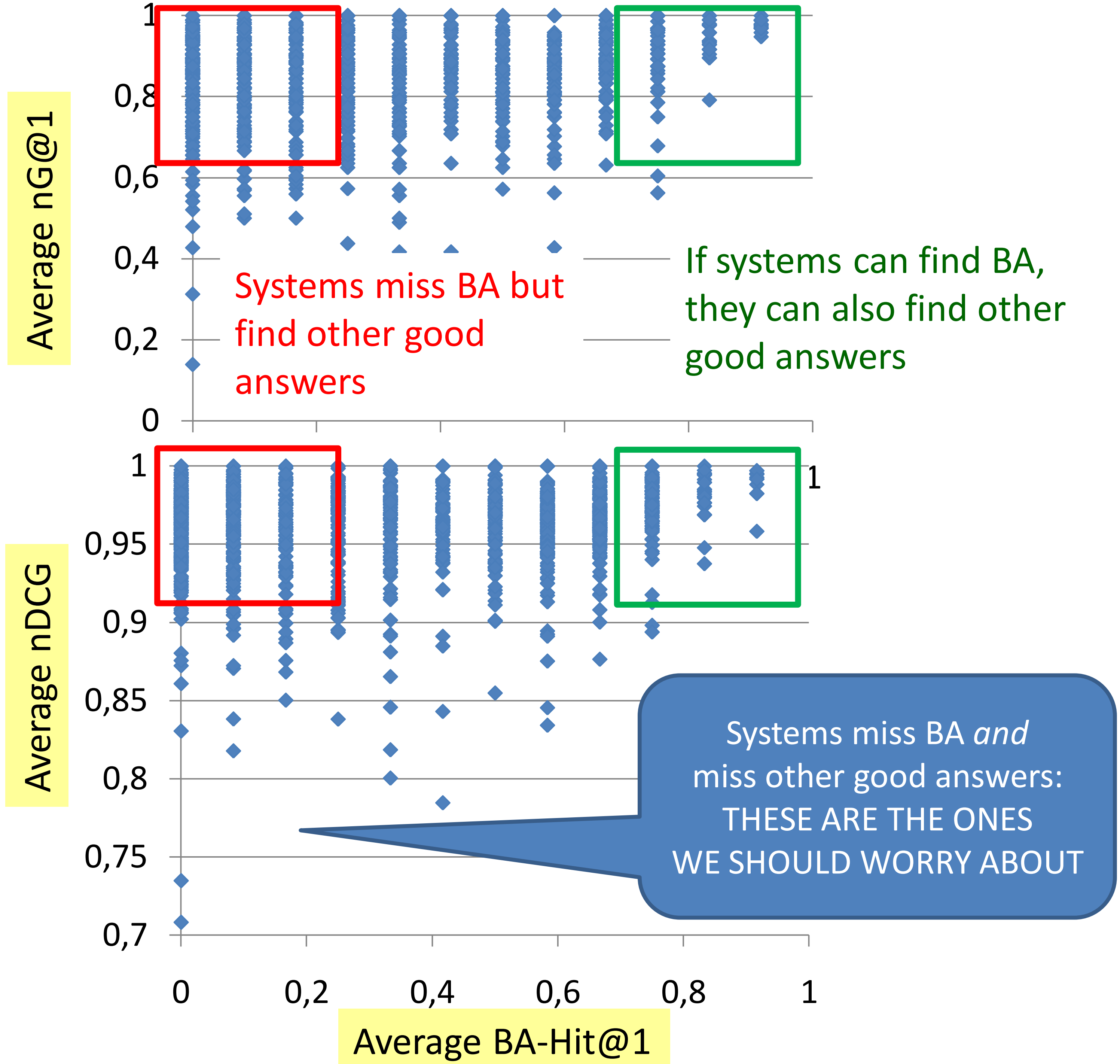
#System pairs significantly different with nDCG=100

#System pairs significantly different with Q-measure=100

Two-sided sign test results

Our method can detect many substantial differences that would have been overlooked by BA-based evaluation.

Finding hard Qs



Easy/hard Qs and categories

- "Easy Qs": Top 500 Qs in average performance
- "Medium Qs": Middle 500 Qs
- "Hard Qs": Bottom 500 Qs

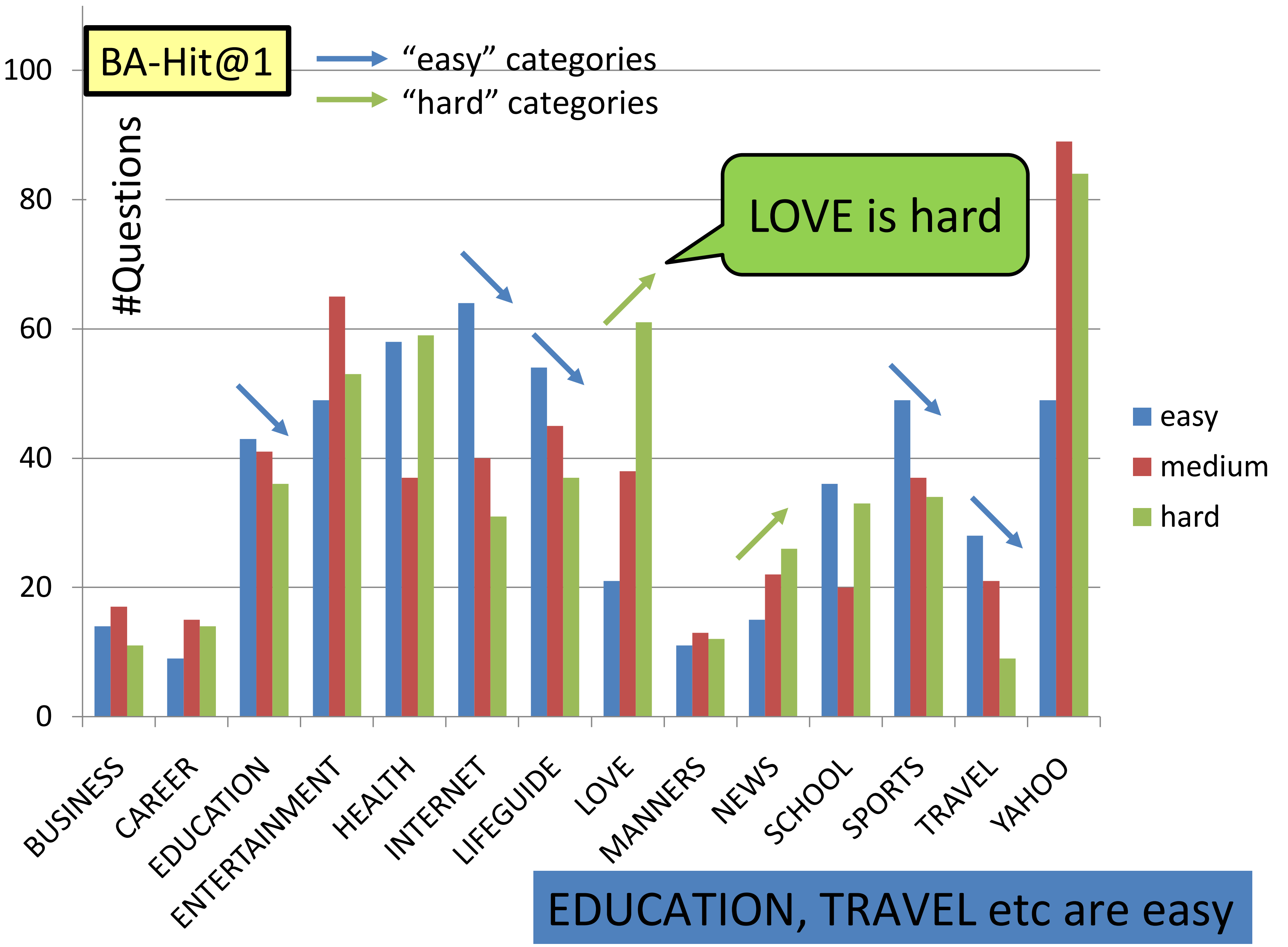
Defined for each metric

- "Easy Q categories"

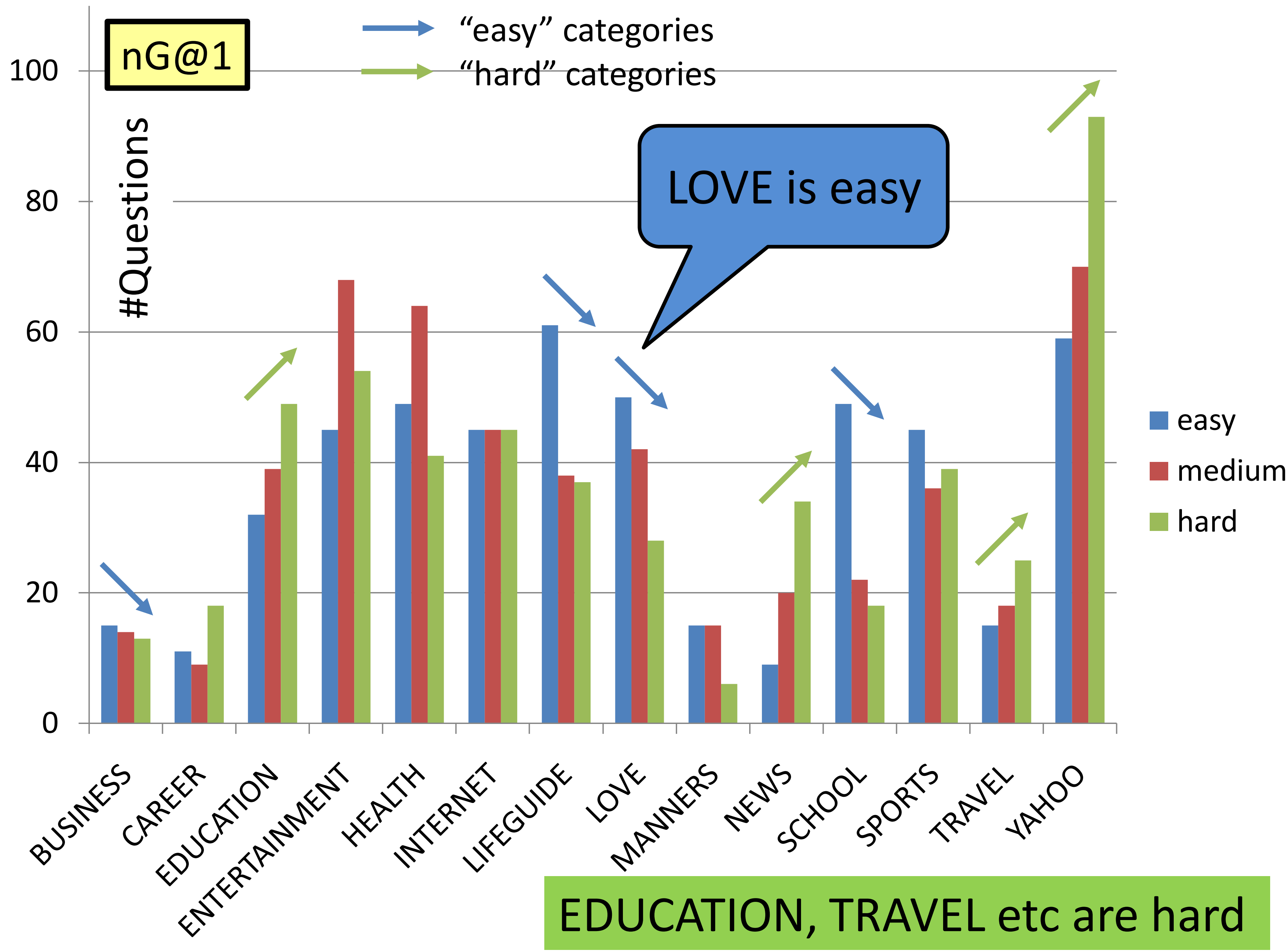
#Easy Qs > #Medium Qs > #Hard Qs

- "Hard Q categories"

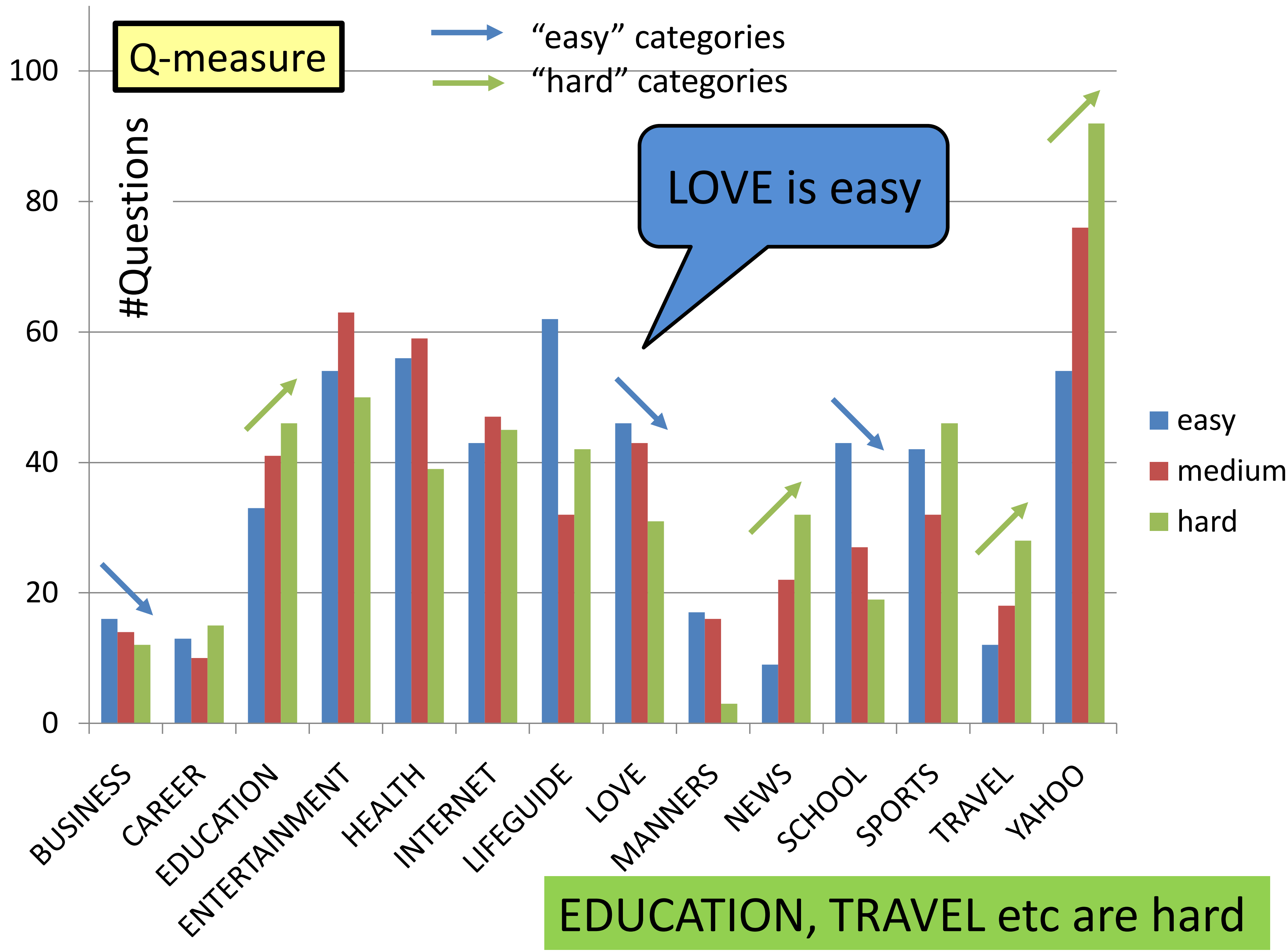
#Easy Qs < #Medium Qs < #Hard Qs



BA-Hit@1
"easy" categories
"hard" categories
#Questions
LOVE is hard
EDUCATION, TRAVEL etc are easy
easy
medium
hard
0
20
40
60
80
100
BUSINESS
CAREER
EDUCATION
ENTERTAINMENT
HEALTH
INTERNET
LIFEGUIDE
LOVE
MANNERS
NEWS
SCHOOL
SPORTS
TRAVEL
YAHOO



nG@1
"easy" categories
"hard" categories
#Questions
100
80
60
40
20
0
LOVE is easy
easy
medium
hard
BUSINESS
CAREER
EDUCATION
ENTERTAINMENT
HEALTH
INTERNET
LIFEGUIDE
LOVE
MANNERS
NEWS
SCHOOL
SPORTS
TRAVEL
YAHOO
EDUCATION, TRAVEL etc are hard


Q-measure
"easy" categories
"hard" categories
#Questions
LOVE is easy
EDUCATION, TRAVEL etc are hard
easy
medium
hard
0
20
40
60
80
100
BUSINESS
CAREER
EDUCATION
ENTERTAINMENT
HEALTH
INTERNET
LIFEGUIDE
LOVE
MANNERS
NEWS
SCHOOL
SPORTS
TRAVEL
YAHOO

[Aikawa/Sakai/Yamana WebDB Forum 2010]

- They manually classified 1500 Qs into Subjective and Objective using two assessors.
- If we look at their results by Q category:

category	SUB	OBJ	Conflict	total
LOVE	117	0	3	120
EDUCATION	24	91	5	120
TRAVEL	13	42	3	58
:				
all	683	749	68	1500

LOVE subjective; EDUCATION/TRAVEL objective

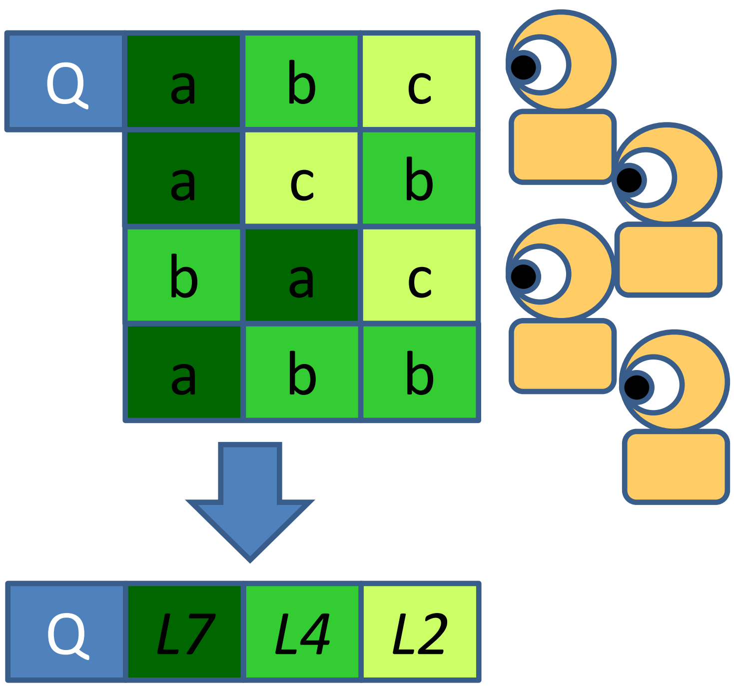
Question hardness, subjectivity and reusability

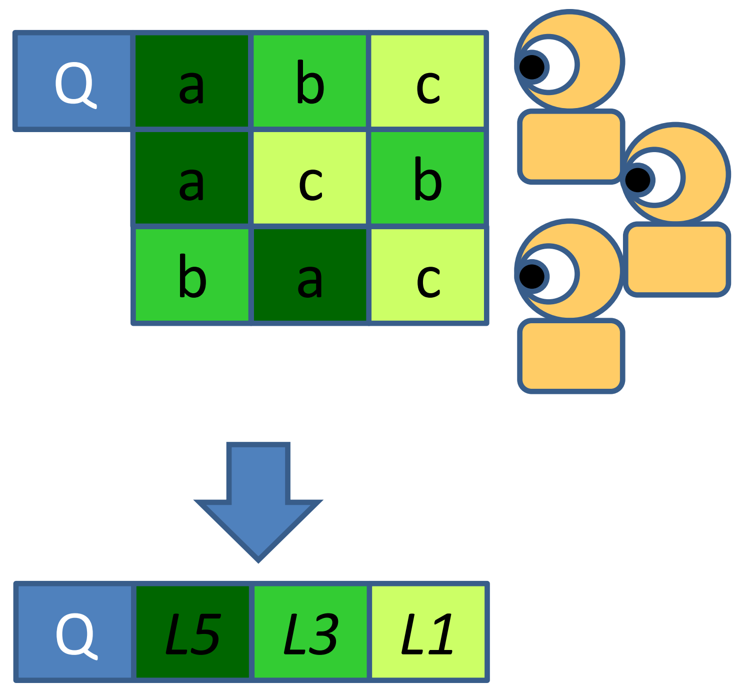
Q category	BA eval	Graded eval	Sub/obj	reusability
LOVE	hard	easy	subjective	maybe
EDUCATION	easy	hard	objective	high
TRAVEL	easy	hard	objective	high

Graded relevance evaluation suggests that systems should improve on EDUCATION, TRAVEL etc since they currently cannot find good answers that are not the BA. These categories are *objective* and important for reuse.

More details in the paper!

- How to normalise nDCG more properly for the task of ranking all answers
- Leave-One-Judge-Out experiments





TALK OUTLINE

1. MOTIVATION
2. PROPOSAL
3. EXPERIMENTS
4. CONCLUSIONS

Conclusions

- Our method can detect many substantial performance differences that would have been overlooked by BA-based evaluation.
- Our method can better identify hard questions (those that are handled poorly by many systems and therefore deserve investigation) compared to BA-based evaluation.

The cost of assessments is worthwhile!
(But what is the optimal number of assessors?
Assessor quality probably more important.)

COMMERCIAL

NTCIR-9 (final meeting: Dec 6-9, 2011, Tokyo)

- INTENT/1CLICK
- Interactive Visual Exploration
- Recognizing Inference in Text
- CrossLingual Link Discovery
- Geotemporal Information Retrieval
- Patent Machine Translation
- IR for Spoken Documents

Open to everyone!