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
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

(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

CQA site data

Previous BA-based evaluation

Automatically extract BA and treat it as the only right answer

Binary relevance data

Proposed

Multiple assessors assess all answers (a/b/c)

Consolidate multiple assessments

Graded relevance data

Consolidate multiple assessments

Previous BA-based evaluation

Automatically extract BA and treat it as the only right answer

Binary relevance data
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

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

Grades by 4 judges
Mapped to 9-point relevance
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]
nDCG vs Q-measure

nDCG = \frac{\sum_{r=1}^{l} g(r) / \log(r + 1)}{\sum_{r=1}^{l} g^*(r) / \log(r + 1)}

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
Finding good systems

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

Performances of 4 judges vary wildly – using multiple judges makes sense.
Discriminative power [Sakai SIGIR06]

Two-sided sign test results:

1. System pairs significantly different with BA-Hit@1 = 59:
   - System pairs: 37
   - Significant differences: 58

2. System pairs significantly different with nG@1 = 95:
   - System pairs: 47
   - Significant differences: 53

3. System pairs significantly different with nDCG = 100:
   - System pairs: 48
   - Significant differences: 52

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

Systems miss BA but find other good answers

If systems can find BA, they can also find other good answers

Systems miss BA and miss other good answers: THESE ARE THE ONES WE SHOULD WORRY ABOUT
Easy/hard Qs and categories

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

- "Easy Q categories"
  \#Easy Qs > \#Medium Qs > \#Hard Qs

- "Hard Q categories"
  \#Easy Qs < \#Medium Qs < \#Hard Qs

Defined for each metric
BA-Hit@1

#Questions

EDUCATION, TRAVEL etc are easy

LOVE is hard

“easy” categories

“hard” categories

EDUCATION, TRAVEL etc are easy
LOVE is easy

EDUCATION, TRAVEL etc are hard
LOVE is easy

EDUCATION, TRAVEL etc are hard
• They manually classified 1500 Qs into Subjective and Objective using two assessors.
• If we look at their results by Q category:

<table>
<thead>
<tr>
<th>category</th>
<th>SUB</th>
<th>OBJ</th>
<th>Conflict</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOVE</td>
<td>117</td>
<td>0</td>
<td>3</td>
<td>120</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>24</td>
<td>91</td>
<td>5</td>
<td>120</td>
</tr>
<tr>
<td>TRAVEL</td>
<td>13</td>
<td>42</td>
<td>3</td>
<td>58</td>
</tr>
<tr>
<td>all</td>
<td>683</td>
<td>749</td>
<td>68</td>
<td>1500</td>
</tr>
</tbody>
</table>

LOVE subjective; EDUCATION/TRAVEL objective
### Question hardness, subjectivity and reusability

<table>
<thead>
<tr>
<th>Q category</th>
<th>BA eval</th>
<th>Graded eval</th>
<th>Sub/obj</th>
<th>reusability</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOVE</td>
<td>hard</td>
<td>easy</td>
<td>subjective</td>
<td>maybe</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>easy</td>
<td>hard</td>
<td>objective</td>
<td>high</td>
</tr>
<tr>
<td>TRAVEL</td>
<td>easy</td>
<td>hard</td>
<td>objective</td>
<td>high</td>
</tr>
</tbody>
</table>

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!