Microsoft Research Asia With Redmond at the NTCIR-8 Community QA Pilot Task

Young-In Song, Jing Liu, Tetsuya Sakai, Xin-Jing Wang, Guwen Feng, Yunbo Cao, Hisami Suzuki and Chin-Yew Lin
Microsoft Research Asia & Microsoft Research
Overview

• Best quality answer finding task in NTCIR
  – For a given QA thread consisting of one question $q$ and its answers $a_1, \ldots, a_n$ ($n \geq 1$), rank answers according to their quality for $q$.

  – Can be regarded as a statistical learning problem on a preference to the best quality answer in a QA thread
    • An answer is represented as a feature vector
    • A statistical model is trained by regarding the best answer selected by a user as a good quality answer
Four aspects in feature selection

**Relevance to question**
- Obviously, quality of an answer should be defined in the context of a question.

**Authority and expertise of answerer**
- A highly authoritative users with expert knowledge on a question domain will be more likely to give a good quality answer.

**Informativeness of answer**
- A good quality answer generally contains rich and detail information for a question.

**Discourse and modality**
- A discourse structure of QA threads (e.g., a position of an answer) or modality of an answer (kindness) can be an effective evidence.
<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>Examined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevance</td>
<td>Unigram LM relevance score</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Graph-based relevance score</td>
<td>Examined in other task</td>
</tr>
<tr>
<td>Authority &amp; Expertise</td>
<td>Number of best answers posted by a user</td>
<td>Examined in other media</td>
</tr>
<tr>
<td></td>
<td>Success rate of a user to post best answers</td>
<td>Newly examined</td>
</tr>
<tr>
<td></td>
<td>Likelihood to be a winner</td>
<td>Newly examined</td>
</tr>
<tr>
<td></td>
<td>Relevance of question to user’s expertise</td>
<td>Newly examined</td>
</tr>
<tr>
<td>Informativeness</td>
<td>Length of an answer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Existence of URL address in an answer</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lexical centrality of an answer in a thread</td>
<td>Newly examined</td>
</tr>
<tr>
<td>Discourse</td>
<td>Position of answers</td>
<td>With new aspect</td>
</tr>
<tr>
<td></td>
<td>Use of negative words</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Agreement relation between Q and A</td>
<td></td>
</tr>
</tbody>
</table>
New Features

• Likelihood to be Winner (LW)
  – Answer graph based on winner-loser relations
  – Each QA thread is viewed as a competition, in which the winner is the answerer posting BA, and others are losers
  – A directed graph can be constructed from QA threads by linking losers to winners

\[ P_t(u_i) = \lambda \cdot P_0(u_i) + (1 - \lambda) \cdot \left( \sum_{\forall u_j} T(u_j \rightarrow u_i) P_{t-1}(u_j) \right) \]

where

\[ P_0(u_i) = \frac{C(BA; u_i)}{\max_{\forall u} C(BA; u)} \]

\[ T(u_j \rightarrow u_i) = \begin{cases} \frac{\text{# of questions } u_j \text{ wins}}{\text{# of questions that } u_j \text{ participate}} & \text{if } u_j = u_i, \\ \frac{\text{# of questions that } u_j \text{ participate}}{\text{# of questions } u_i \text{ wins } u_j} & \text{else} \end{cases} \]
New Features

• User Expertise LM Score (UE)
  – If a question is well matched an answerer’s knowledge, there will be a higher probability that quality of the answer from the answerer is good
    • Build expertise language model from user’s answers
    • Estimate a probability generating a question from one’s expertise model

• Lexical centrality of an answer in a thread (LEX)
  – In terms of informativeness, the best quality answer is the best summary of QA thread
  – Use the possibility of an answer to be a good summary as a feature
    • LexRank approach are applied [Erkan et al, 2004]
New Features

• Position of answers (PA)
  – Top contributors in CQA community have a tendency to answer questions only if necessary [Nam et al, 2009]
    • If there is sufficiently good answer, they will skip the thread
    • The lastly posted answer is more likely to be better quality answer

\[ PA(a_i) = \frac{1}{|T_j| - Pos(a_i)} \]

![Graph showing the probability of being the best answer against normalized answer position](image)
Models

• Classification vs. Pairwise learning
  – Best quality answer finding task can be formulated as a Binary classification task
    • Assuming BAs as good quality answer (positive) and Non-BAs as bad quality answer (negative)
    • Too many false negatives: Some of non-BAs are actually good quality answers
  – Advantage of using pairwise learning approach
    • The assumption is relaxed: ‘BA’ is better than non-BA
      – False negative only happens when non-BA is better than BA in quality
    • SVM rank is used as our default model in the experiments
Models

• Analogical Model [Wang, 2009]
  – Two similar questions may share similar good quality answers
    • By utilizing previously-posted QA threads similar to a new question, a better answer quality evaluation would be possible
  – One problem on test data configuration
    • All questions in NTCIR test data are found in the training data
    • Under this setting, the analogical model will take unrealistic advantages
      – Always, it has a chance to optimize model parameters based on ‘correct best answers’
Run Configuration

Run 3 is a run extensively using authority and expertise features, and Run 4 is a run mainly to examine relatively new features.

Run 5 is a run to test analogical model with the basic feature set (same to Run 1).

Run 1 is the simplest system designed with minimum number of features.
Run 2 represents the most effective system using all features effective in our preliminary experiments with BA.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Run 1</th>
<th>Run 2</th>
<th>Run 3</th>
<th>Run 4</th>
<th>Run 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relevance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LMRS$^3$</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>GRS</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Authority and Expertise</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NBA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>PS</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>LW</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>UE</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>Informativeness</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NLA</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>URL</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>LEX+NLA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>LEX+PS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>Discourse and Modality</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>NW</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>AR</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th></th>
<th>BA-Hit@1</th>
<th>GA-Hit@1</th>
<th>GA-nG@1</th>
<th>GA-nDCG</th>
<th>GA-Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run 2</td>
<td>0.4980 (ΔB2, ΔR1)</td>
<td>0.9967 (ΔB2)</td>
<td><strong>0.9211</strong> (ΔB2, ΔR1)</td>
<td>0.9747 (ΔB2, ΔR1)</td>
<td><strong>0.9690</strong> (ΔB2, ΔR1)</td>
</tr>
<tr>
<td>Run 1</td>
<td>0.4980 (ΔB2)</td>
<td>0.9967 (ΔB2)</td>
<td>0.9203 (ΔB2)</td>
<td>0.9741 (ΔB2)</td>
<td>0.9682 (ΔB2)</td>
</tr>
<tr>
<td>Run 4</td>
<td>0.4847</td>
<td>0.9973 (ΔB2, ΔR1)</td>
<td>0.9202 (ΔB2)</td>
<td>0.9745 (ΔB2, ΔR1)</td>
<td>0.9688 (ΔB2, ΔR1)</td>
</tr>
<tr>
<td>Baseline-2 (Length only)</td>
<td><strong>0.4847</strong></td>
<td>0.9953</td>
<td><strong>0.9170</strong></td>
<td>0.9735</td>
<td><strong>0.9680</strong></td>
</tr>
<tr>
<td>Run 3</td>
<td>0.4813</td>
<td>0.9960 (ΔB2)</td>
<td>0.8956</td>
<td>0.9679</td>
<td>0.9609</td>
</tr>
<tr>
<td>Run 5</td>
<td><strong>0.7773</strong> (ΔB2, ΔR1)</td>
<td>0.9987 (ΔB2, ΔR1)</td>
<td>0.8863</td>
<td>0.9604</td>
<td>0.9499</td>
</tr>
<tr>
<td>Baseline-3 (Posting Time)</td>
<td>0.3820</td>
<td>0.9940</td>
<td>0.8213</td>
<td>0.9460</td>
<td>0.9359</td>
</tr>
<tr>
<td>Baseline-1 (Random)</td>
<td>0.2713</td>
<td>0.9920</td>
<td>0.7751</td>
<td>0.9311</td>
<td>0.9169</td>
</tr>
</tbody>
</table>

### Askers (best answers) vs. our system?

<table>
<thead>
<tr>
<th></th>
<th>GA-nG@1</th>
<th>L3-Hit@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run 2</td>
<td><strong>0.9211</strong></td>
<td><strong>0.8054</strong></td>
</tr>
<tr>
<td>BA as top 1 rank</td>
<td>0.8900</td>
<td>0.7315</td>
</tr>
</tbody>
</table>
Observations

• The best answer selected by an asker is not only the best answer and often it is not really the best answer

• Length is a very powerful feature in best quality answer finding
  – The improvements by other features were only marginal
  – The Ga-nG@1 and GA-nDCG score of length-based ranking: 0.9170 / 0.9735

• How to train a model better based on noisy and partial positive examples?