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

Features

Type	Name	
Relevance	Unigram LM relevance score	
	Graph-based relevance score	Examined in other task
Authority & Expertise	Number of best answers posted by a user	Examined in other media
	Success rate of a user to post best answers	Newly examined
	Likelihood to be a winner	Newly examined
	Relevance of question to user's expertise	Newly examined
Informativeness	Length of an answer	
	Existence of URL address in an answer	
	Lexical centrality of an answer in a thread	Newly examined
Discourse	Position of answers	With new aspect
	Use of negative words	
	Agreement relation between Q and A	

New Features

- Likelihood to b
  - Answer graph
    - Each QA thre
      the answerer
    - A directed gr
      losers to win

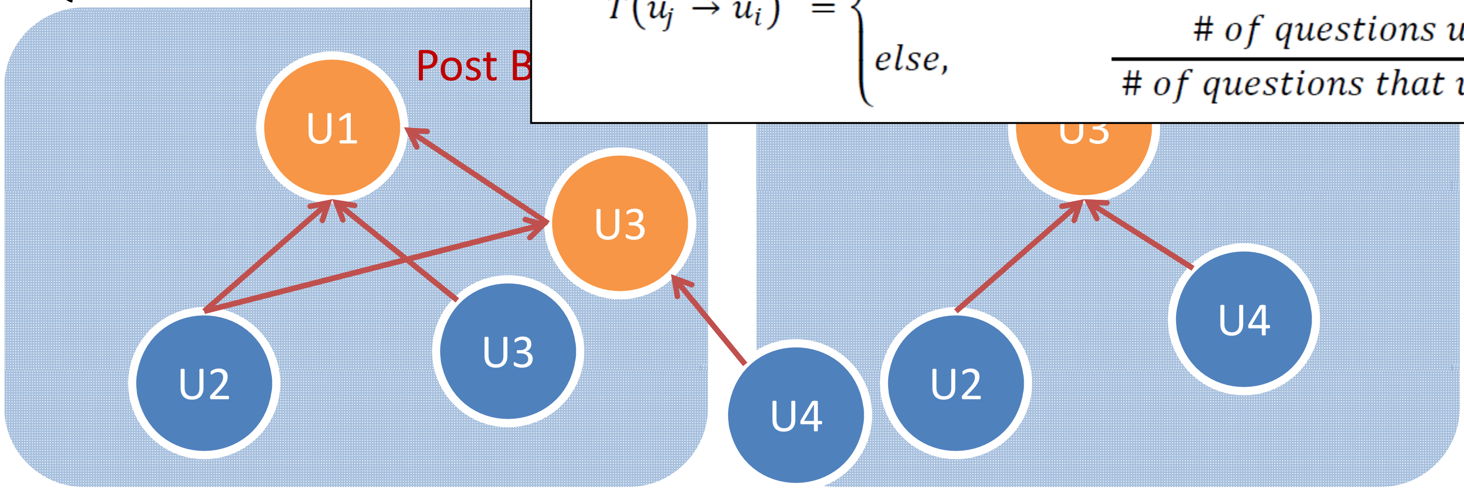
$$P_t(u_i) = \lambda \cdot P_0(u_i) + (1-\lambda) \cdot \left(\sum_{\forall u_j} T(u_j \rightarrow u_i) P_{t-1}(u_j) \right)$$

where

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

$$T(u_j \rightarrow u_i) = \begin{cases} if\ u_j = u_i, & \dfrac{\#\ of\ questions\ u_j\ wins}{\#\ of\ questions\ that\ u_j\ participate} \\ else, & \dfrac{\#\ of\ questions\ u_i\ wins\ u_j}{\#\ of\ questions\ that\ u_j\ participate} \end{cases}$$

QA thread 1

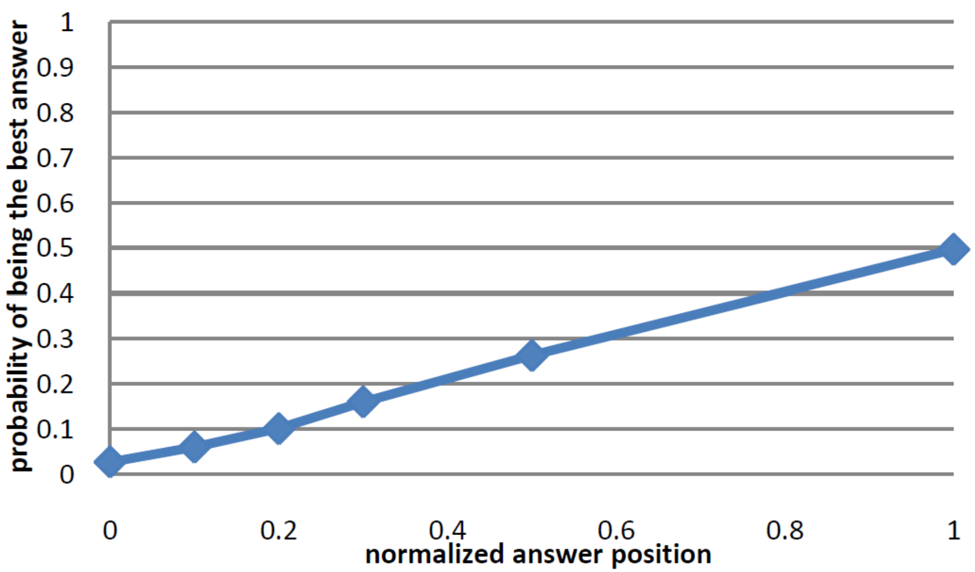


New Features

- 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



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

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

Table 1: Feature configurations of

	Feature	Run 1	Run 2	Run 3	Run 4	Run 5
Relevance	LMRS[3]					
	GRS	√	√			√
Authority and Expertise	NBA			√		
	PS	√	√	√		√
	LW			√	√	
	UE		√	√	√	
Informativeness	NLA	√	√			√
	URL	√	√	√	√	√
	LEX+NLA				√	
	LEX+PS			√		
Discourse and Modality	PA	√	√	√	√	√
	NW		√	√	√	
	AR		√	√	√	

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 BAs

Results

	BA-Hit@1	GA-Hit@1	GA-nG@1	GA-nDCG	GA-Q
Run 2	0.4980 (Δ_{B2}, Δ_{R1})	0.9967 (Δ_{B2})	**0.9211 (Δ_{B2}, Δ_{R1})**	**0.9747 (Δ_{B2}, Δ_{R1})**	**0.9690 (Δ_{B2}, Δ_{R1})**
Run 1	0.4980 (Δ_{B2})	0.9967 (Δ_{B2})	0.9203 (Δ_{B2})	0.9741 (Δ_{B2})	0.9682 (Δ_{B2})
Run 4	0.4847	0.9973 (Δ_{B2}, Δ_{R1})	0.9202 (Δ_{B2})	0.9745 (Δ_{B2}, Δ_{R1})	0.9688 (Δ_{B2}, Δ_{R1})
Baseline-2 (Length only)	0.4847	0.9953	0.9170	0.9735	0.9680
Run 3	0.4813	0.9960 (Δ_{B2})	0.8956	0.9679	0.9609
Run 5	**0.7773 (Δ_{B2}, Δ_{R1})**	**0.9987 (Δ_{B2}, Δ_{R1})**	0.8863	0.9604	0.9499
Baseline-3 (Posting Time)	0.3820	0.9940	0.8213	0.9460	0.9359
Baseline-1 (Random)	0.2713	0.9920	0.7751	0.9311	0.9169

Askers (best answers) vs. our system?

	GA-nG@1	L3-Hit@1
Run 2	**0.9211**	**0.8054**
BA as top 1 rank	0.8900	0.7315

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?