QUESTION ROUTING IN COMMUNITY QUESTION ANSWERING SERVICES

Haritha C. V[1], Maya Mohan[2], and Sruthy Manmadhan[3]
[1]M.Tech First Year Student, [2,3]Assistant Professor
Department of Computer Science and Engineering,
N.S.S College of Engineering, Palakkad
Email: [1]hbharithababu@gmail.com, [2]mayajeevan@gmail.com, [3]sruthym.88@gmail.com

***Abstract*—With the development of Web 2.0, community question answering services like Yahoo! Answers, have attracted great attention from both academia and industry. In CQAS, anyone can ask and answer questions on any topic. As answers are usually explicitly provided by humans, they can be helpful in answering real world questions. There are numerous problems in this area: One of the main problem arise here is regarding the case of question retrieval. Word ambiguity and word mismatch problem bring new challenges in question retrieval. This can be corrected by taking advantage of potentially rich semantic information drawn from other languages along with non-negative matrix factorization. There are certain scenarios where the efficiency and answer quality of services may not up to the level of satisfaction. Question routing has been proposed to solve this by routing new question to eligible answer. This is addressed by "Multi objective learning to rank approach" which simultaneously optimizes answering possibility and answer quality of routed users. Another challenging problem is that questions may remain unanswered in CQA sites. To alleviate such problem there is a novel scheme to rank answer candidates of similar questions via pairwise comparison and choose best among them to answer new question. The same question starvation problem can also be address by means of coupled mutual reinforcemet scheme which uses LambdaMART. Answer selection in community question answering is another challenging issue in natural language processing. It is possible to handle this problem using an attentive deep neural network architecture so as to learn deterministic information for answer selection.**

Keywords: CQAS, Matrix factorization, Multi-objective learning to rank method, Pairwise comparison scheme, LambdaMART, Attentive neural network

I. Introduction

Search engines are the most commonly used application in our daily life to find the answers to questions. But, they require users to know the effective search keywords, without which users needs to spend an extremely long time in searching for answers. Some of the user questions are typically personal, heterogeneous, and open-ended, search engines are not intelligent enough to find a single web page that can directly answer such questions. Since real humans are believed to understand and answer better than a machine, Community Question Answering Services (CQAS) provide a platform to allow people to post questions and answer questions posted by others. In a CQAS, a question is open for receiving answers during a certain time period. An asker can select the best answer for his/her question along with a rating in a given range.

Thus, CQAS, is the fastest-growing user-generated-content (UGC) portals, and has risen as an enormous market, so to speak, for the fulfillment of complex information needs. CQAS enables users to ask/answer questions and search through the archived historical question-answer (QA) pairs. Compared to the traditional factual QA, such as "who is the prime minister of the India in 2017", which can be answered by simply extracting named entities or paragraphs from documents, CQA have made substantial headway in answering complex questions, such as reasoning, open-ended, and advice-seeking questions. CQAS is thus quite open and has little restrictions, if any, on who can post and who can answer a question. Some of the popular sites are as follows:

- **Yahoo! Answers**: It is basically a Community question answering site from yahoo and it allows a person to submit queries that needs to be answered and answer questions asked by other users.
- **Quora**: Quora is a community question-and-answer site in which the questions are asked, answered, edited as well as organized by the group of users.
- **Stack Overflow**: This community question answering website act as a platform for users to ask their question and answer questions of others , and, through active participation, they can also post votes for questions as well as answers up or down and edit questions and answers similar to that of wiki. Users of Stack Overflow can earn reputation points.
- **HealthTap**: HealthTap is an interactive health company invented to reinvent worldwide the way people take care of their health. It is applicable to both for patients and for doctors, is available across a range of platforms and devices.

II. Related Work

A. Answer Selection in CQA using Attentive Neural Network

The major task of answer selection in CQA lies on extracting useful QA pairs from multiple CQA threads. But, the main problem is how to bridge the semantic gap between Question Answer pairs. The problem can be addressed by using an attentive neural network [1] that learn the deterministic information for answer selection from a global perspective. The architecture is made by several network components including Convolutional Neural Networks (CNN), attention based LSTM and Conditional

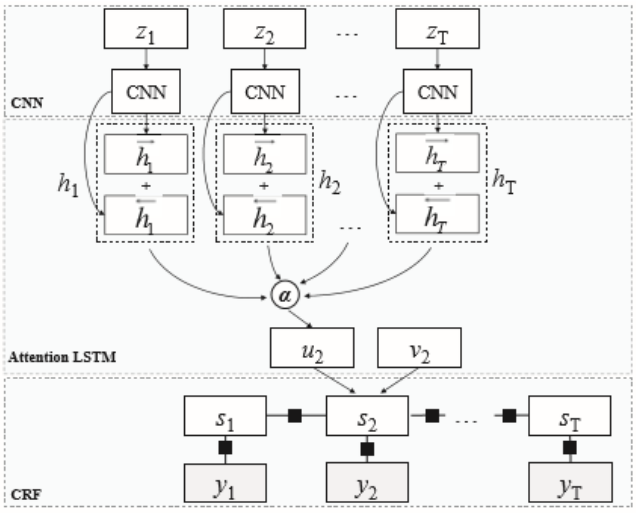

Figure 1. The overview of the proposed attentive neural architectures [1]

Random Fields (CRF).

Figure 1 shows the architecture of the proposed system. The main components include CNN, attention LSTM and CRF. CNN act as the first encoder that extracts features from each input sentence in each time step and compresses them into a fixed length vector. Since in CQAS most of the question are short, one could take each question as a sentence [2] and apply word embedding, multiple convolution and max-pooling operations. Then, Attention LSTM further encodes the compressed vectors and learns long term dependencies [3] from the sequential steps. It is constituted by a bidirectional LSTM layer (h_t) followed by a soft attention layer (α), which can capture the correlations along the whole sequence in a large part. Before information propagated to the next layer, A-ARC enables the addition of external elements (v_t) so that it is possible to add other features beyond the learned dependencies (u_t). So, Simply speaking an LSTM layer can preserve information from early time step by controlling gates and memory cells. The last component is CRF, which aims to generate final predictions considering both the encoded representation (s_t) as well as label transitions. That is, Given an observation sequence s = s_1, s_2 ,..., s_T CRF models the generating probability of the entire label sequence y = y_1 ,y_2 ,..., y_T by using the discriminative probabilities to y_i given s_i and the transition probabilities between adjacent labels.

B. Multi-objective Optimization Approach to Question Routing in CQA services

Even though CQA services benefit users using human-generated answers, main problem has arisen over the efficiency as well as answer quality of these services. In many situation the askerer may have to wait for several days to receive a response, and also the people may have to browse many questions before finding one they want to answer; At the same time, many answers are less in quality, from which the askers can get very little help. Question routing is the best solution for these problems that route new questions to eligible answerers. This is addressed by formulating question routing as a multi-objective ranking problem [4] that is, for a new question which is being posted, rank a set of users and route this question to the users who has high ranks, along with both high answering possibility and high answer quality at the same time. This approach is termed as a multi-objective learning-to-rank approach for question routing, which is referred to as MLQR.

The main idea of MLQR is: Training a multi-objective ranking model that map a list of user features to a list of scores for ranking these users. The users interests are captured by tagword topic model (TTM). Latent Dirchlet Allocation [5] can not be used here since the questions are very short. TTM usually combines a tag with a word to form tag-word co-occurrences and aggregate them on the corpus level to relieve the data sparsity problem of questions. TTM can also build connections between user history posts and the given question. MLQR is performed in two phases. Training and ranking phases. "Training phase", make use of existing questions and using this train a tagword topic model (TTM), which is used to derive topical representations of new questions. Based on TTM, it is possible to capture features like user interest, for a specific question at both platform level and thread level. The platform-level features summarize the information of a user from his history of posts, while the thread-level features learn various ones performance by measuring his competition with other users. Later, Using these features generate instances for training, where an instance consists of users under a question represented by their features along with two relevance labels, Finally, use an extended learning-to-rank algorithm based on LambdaMART [6] to train a multi-objective ranking model, which simultaneously consider both answering possibility and answer quality. In the second phase of MLQR that is, "ranking phase", given a new question and a set of users, first extract user features, calculate their scores by the trained model, and rank the users in descending order of their obtained scores. Finally perform routing task by selecting users with high scores so as to accomplish the task of question routing.

Table 1 illustrates some of the basic features extracted during training phase and used subsequently.

TABLE I
FEATURE SUMMARY [4]

Feature	Description
Topical Similarity	Similarity between topic distribution
Answer ratio	Number of answers
Number of questions	Number of questions posted by the user
Number of answers	Number of answers posted by the user
Number of posts	Number of questions posted + answers posted
Total votes	Total votes the user has received
WINlength and LOSSlength	Length of the answer detail is selected

C. Multilingual Translation Representation to Question Retrieval in CQA via NMF

Question retrieval is a task of finding similar questions to that of a queried question and using historical questions to answer the same [7]. But, is a challenging task due to the word ambiguity and word mismatch between the queried questions and the historical questions. Word ambiguity causes the retrieval models to retrieve many historical questions that do not match the users objective while word mismatch means that the queried questions may contain words that are different from, but having same meaning to, the words in the relevant historical question. Translated representation [8] is the better solution to this which will alternatively enrich the original questions with the words from foreign languages. The idea of improving question retrieval with statistical machine translation is based on the following two observations: (1) Contextual information is utilized during the translation. (2) Multiple words that have similar meanings in one language may be translated into an unique word in a foreign language. But there are two problems with this enrichment: (1) Enrichment makes the question representation even more sparse; (2) Introduction of noise, which can harm the performance of question retrieval. To solve these two problems, here is a leverage statistical machine translation to improve question retrieval via non-negative matrix factorization [9].

According to figure 2, Let L = l_1,l_2,...,l_P denote the language set, where P is the number of languages, l_1 denotes the original language while l_2 to l_P are the foreign languages. Let $D_1 = d_1^{(1)}$, $d_2^{(1)}$,...,$d_N^{(1)}$ be the set of historical question collection in original language, where N is the number of historical questions in D_1 with vocabulary size M_1. First translate original historical question from language l_1 onto any other languages l_p by Google Translate. Thus, can obtain D_2,...,D_P in different languages, and M_p is the vocabulary size of D_p. Then enrich the original question representation by adding the translated words from language l_2 to l_P. The major problems in enrichment can be avoided by combing Statistical machine learning with matrix factorization. NMF [10] aims to find two non-negative matrices which combines to form the original matrix again, on multiplication. Thus, NMF is used to induce the reduced representation V_p of D_p. U_p ensures that the reduced representation for terms does not overfit. Finally, the relevance score between the queried question q_1 and the historical question d_1 in the reduced space is, then, calculated as the cosine similarity between v_{q1} and v_{d1} as in equation (1).

$$s(q_1, d_1) = \frac{< v_{q1}, v_{d1} >}{||v_{q1}||^2 * ||v_{d2}||^2} \tag{1}$$

Finally relevance function is found by BM25.

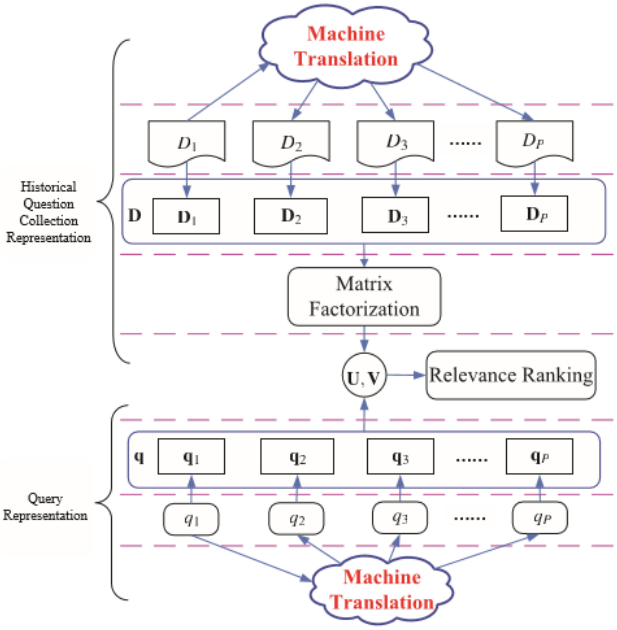

Figure 2. Proposed Architecture [9]

D. Data-driven Answer Selection in Community QA Systems

Question starvation is the major problem that is being faced in CQA forums. It is a condition in which a question remain unanswered for a long time. So, the problem of question starvation is addressed in a different manner, that is by using Pairwise Learning to rANk model, nicknamed PLANE [11], which can rank answer candidates from the relevant question pool. PLANE model, consist of two components: offline learning and online search. During the offline learning stage, it automatically establishes the positive, negative, and neutral training samples with respect to preference pairs. The PLANE model is made by jointly training with these three kinds of training samples. When it comes to the online search, for a given question, pair the question with each of the answer candidates in the candidate pool, and fit them into the trained PLANE model to estimate their matching scores.

1) Offline Learning: Let a_i^j be the j-th answer of the i-th question q_i, and a_i^0 is the best answer. According to the first and third observations in figure 3, it is possible to make equation (2) and equation (3)

$$(q_i, a_i^0) > (q_i, a_i^j), j \neq (0) \tag{2}$$

$$(q_i, a_i^j) > (q_i, a_k^t), i \neq (k) \tag{3}$$

where $>$ denotes a preference relationship.

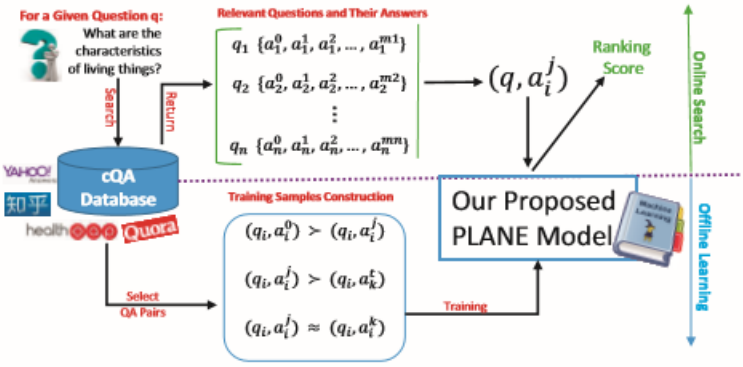

Figure 3. PLANE Model [11]

Let x = $x^{(1)}$ - $x^{(2)}$, where $x^{(1)}$ and $x^{(2)}$ respectively denote the D-dimensional feature vectors of the first and second QA pairs in each comparison. Meanwhile, y be the preference relationship of x, which satisfies the conditions indicated by equation (4) and equation (5):

$$y = 1 \ \ when \ \ x^{(1)} > x^{(2)} \tag{4}$$

$$y = -1 \ \ when \ \ x^{(2)} > x^{(1)} \tag{5}$$

Using this observation the training set can be constructed by: X=(x_i , y_i). According to the second observation in figure 3, there does not exist a preference relationship among the non best answers of a specific question. So, there exists neutral preference relationship between the first and second QA pairs. Using this create another training set with neutral preference U =(u_j,0). The combined effect of X and U gives pairwise learning to rank model.

2) Online Search: As a first stage search the repositories to find similar questions to the newly posted question. This work can done by using Cosine similarity, syntactic tree [12] etc but here preferred Apache Lucene-based k-NN strategy that find the top k similar questions. So, it is possible to create an answer candidate pool by gathering all the answers associated to the k returned questions. Then pair the given question with each of the answers in the pool. Following that, utilize this model to generate an answer ranking list by pairwise comparison.

E. Question Quality Analysis with Coupled Mutual Reinforcement

A large portion of questions remain unanswered in CQA systems, this phenomenon is referred to as question starvation [13]. To solve this problem, the atmost importance lies on improving the Question Quality (QQ) in order to attract more attention to a question from users. The factors influencing QQ, is determined by making use of 4 features such as: question, asker, category and answer related features. For the correlation analysis, it is important to consider the mutual interactions between multiple classes of features. But, traditional classification algorithms cannot consider such mutual interactions. So, here uses a graph-based Coupled Semi-Supervised Mutual Reinforcement algorithm (CSMRLP)[14]. Table II shows important features that are extracted. Prior asker-answerer interaction [15] is important among them. To identify the features that influence the QQ, and determine the degree of influence of each feature on the QQ, computed the information gain (IG) of each feature in Table II. Then, ranked the features based on their IGs and it is easy to find important feature that effect QQ.

TABLE II
FEATURES AND DESCRIPTION [11]

Feature	Description
Subject length	Number of words in the question
Content length	Number of words in the content
Category matching	How well the question matches with category
Capital error	Number of capital errors
Askers points	Points earned from the beginning
Question resolved	Question resolved in the past
Prior asker-answerer interaction	Number of prior interactions
Best answer length	The length of the best answers

F. Prediction Model

Use a bipartite graph G_{ij} = (N_{ij} , E_{ij}) to model the relation between asker expertise and QQ. N_{ij} is the set of nodes and E_{ij} is the set of undirected edges. N_{ij} consists of two types of nodes that correspond to the questions and the askers, respectively. There is an edge between a question node and an asker node if the question is asked by the asker.

1) Finding QQ and askers asking expertise. : First propagates user expertise by finding askers asking expertise from their neighbors and their QQ. Second, the propagates QQ. Then it clamps the labeled data of askers asking expertise and questions qualities. After repeating a certain number of times, the algorithm can estimate all questions qualities and askers asking expertise.

2) Finding QQ and the reputation of categories. : First propagates category reputation by estimating the reputation of categories from their neighbors and their questions qualities. Second, the propagates QQ. Then it clamps the labeled data of reputation of categories and questions qualities. After repeating a certain number of times, the algorithm can estimate all questions qualities and reputation of categories.

3) Finding QQ and AQ. : Finally relation between QQ and AQ can be found by first propagating AQ by estimating answers qualities from their neighbors and their questions qualities Second, the propagates QQ by estimating questions quality from their neighbors and answers qualities. Then it clamps the labeled data of answers qualities and questions qualities. After repeating a certain number of times, the algorithm can estimate all questions qualities and answers qualities.

All the three factors combines to yield QQ.

III. DISCUSSIONS AND FINDINGS

A. Discussion

CQA systems are something that is very much useful to a user. But, at the same time these services faces many challenges but

TABLE III
COMPARISON

Type	Advantages	Disadvantages
Attentive NN	Context modeling	Word embedding
MLQR	Answering possibility & quality	Complex approach
Translations	Noise free & reduce sparsity	Question structure
PLANE	Solve Question starvation	Noise-sensitive.
CSMRLP	Solve question starvation	Answerer-related features

they can be efficiently solved. Semantic similarity between new questions and answers that are present in the historical pool can be checked using an attentive neural network mechanism which works by taking the advantages of Context modeling. But it has a major drawback that it always makes use of word embedding and does not consider any features other than this. Question Routing can be efficiently done in CQAS by a novel MLQR approach that maps users to scores for ranking them. This rank is made based on both answering possibility and answer quality of an answerer to a question. But this is too complicated. The problem of word ambiguity and word mismatch between queried question and historical question can be solved by translation representation that combines both statistical machine learning with non-negative matrix factorization to avoid sparsity and noise. But they are not considering the question structure. Question starvation problem can be addressed in 2 ways. One make use of a novel PLANE model that pairs new question with historical answer and selects the most matching one. But this pairing is sensitive to noise. The second approach makes use of a coupled mutual reinforcement that consider co-relation between QQ and several other features like AQ. But it never make use of answerer related features as one of its concern.

B. Findings

CQAS are more beneficial to user since they are capable of answering user questions which are heterogeneous, specific, and open-ended. The major problem in CQAS is the question starvation problem in which a question posted by a user remains unanswered for a long period of time. The best effective means to address this issue is to make use of pair-wise learning to rank method. But the issue is that it is susceptible to noises. From our knowledge we are familiar that using non-negative matrix factorization, it is possible to reduce the amount of noise which is being generated. Non-negative matrix factorization is a process in which a matrix V of dimension mxn can be decomposed into 2 matrices H and W of dimension mxr and rxn so that W*H = V. So the newly posted question is first factorized as in NMF with respect to historical questions and then collect answers of similar questions and pair new question with these answers. Then by ranking chooses the best answer. Thus, both noise is reduced and question starvation can be minimized.

IV. CONCLUSION

Some user questions are typically personal, heterogeneous, specific and open-ended, search engines are usually not intelligent enough to find a single web page that can directly answer such questions. Community Question Answering Services (CQAS) provide a platform to allow people to post questions and answer questions posted by others. In a CQAS, a question is open for receiving answers during a certain time period. An asker can select the best answer for his/her question along with a rating in a given range. Also, each question has an attribute of tag-of-interests, which represents the number of users interested in this question. But CQAS faces several challenges such as 1). Semantic gap between question answer pairs which can be solved by attentive neural network, 2). Word mismatches and ambiguity problems are addressed using non-negative matrix factorization, 3).Question routing by Multi-objective Optimization Approach and 4) Question starvation by 2 measures namely data driven methods and one via mutual reinforcement.

REFERENCES

[1] Yang Xiang, Qingcai Chen *"Answer Selection in Community Question Answering via Attentive Neural Networks "*, IEEE Transaction Journal, VOL. 14, NO. 8, AUGUST 2016

[2] Y. Kim, (2014), *"Convolutional neural networks for sentence classification"* , arXiv preprint arXiv:1408.5882.

[3]] A.Graves, (2013), *"Generating sequences with recurrent neural networks"*, arXiv preprint arXiv:1308.0850.

[4] Xiang Cheng, Shuguang Zhu, Sen Su, Gang Chen, (2017), *"Multi-objective Optimization Approach for Question Routing in Community Question Answering Sevices"*, IEEE Transaction Journal.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, (2003), *"Latent dirichlet allocation"*, the Journal of machine Learning research, vol. 3, pp.993-1022.

[6] C. J. Burges, (2010), *"From ranknet to lambdarank to lambdamart: An overview"*

[7] Guangyou Zhou, Zhiwen Xie, Tingting He, Jun Zhao and Xiaohua Tony Hu, (2016), *"Learning the Multilingual Translation Representations for Question Retrieval in Community Question Answering via Non-negative Matrix Factorization"* , IEEE Transaction Journal.

[8] H. Duan, Y. Cao, C.-Y. Lin, and Y. Yu, (2008), *"Searching questions by identifying question topic and question focus"*, in ACL, pp. 156-164.

[9] G. Zhou, K. Liu, and J. Zhao, (2012), *"Exploiting bilingual translation for question retrieval in community-based question answering"*, in COLING, pp. 3153-3170

[10] D. D. Lee and H. S. Seung, (2000), *"Algorithms for*

non-negative matrix factorization", in NIPS, pp. 556-562.

[11] Liqiang Nie, Xiaochi Wei, Dongxiang Zhang, Xiang Wang, Zhipeng Gao, and Yi Yang, (2016), *"Data-driven Answer Selection in Community QA System "*, IEEE Transaction Journal.

[12] K. Wang, Z. Ming, and T.-S. Chua, (2009) *"A syntactic tree matching approach to finding similar questions in community-based qa services"*, in Proceedings of SIGIR 09. ACM, pp. 187-194.

[13] Jinwei Liu, Haiying Shen, (2015), *"Question Quality Analysis and Prediction in Community Question Answering Services with Coupled Mutual Reinforcement"* , IEEE Transaction Journal.

[14] B. Li and I. King, (2010), "Routing questions to appropriate answerers in community question answering services", In Proc. of CIKM.

[15] E. Rodrigues and N. Milic-Frayling, (2009), *"Socializing or knowledge sharing? characterizing social intent in community question answering"*, In Proc. of ACM CIKM, pp. 1127-1136, Hong Kong