Quality Matters: Assessing cQA Pair Quality via Transductive Multi-View Learning

Xiaochi Wei[1], **Heyan Huang**[*1], **Liqiang Nie**[2], **Fuli Feng**[3], **Richang Hong**[4], **Tat-Seng Chua**[3]
[1] Beijing ER Center of HLIPCA, School of Computer, Beijing Institute of Technology
[2] School of Computer Science, Shandong University
[3] School of Computing, National University of Singapore
[4] School of Computer and Information, Hefei University of Technology
{wxchi,hhy63}@bit.edu.cn, {nieliqiang,fulifeng93,hongrc.hfut}@gmail.com, chuats@comp.nus.edu.sg

Abstract

Community-based question answering (cQA) sites have become important knowledge sharing platforms, as massive cQA pairs are archived, but the uneven quality of cQA pairs leaves information seekers unsatisfied. Various efforts have been dedicated to predicting the quality of cQA contents. Most of them concatenate different features into single vectors and then feed them into regression models. In fact, the quality of cQA pairs is influenced by different views, and the agreement among them is essential for quality assessment. Besides, the lacking of labeled data significantly hinders the quality prediction performance. Toward this end, we present a transductive multi-view learning model. It is designed to find a latent common space by unifying and preserving information from various views, including question, answer, QA relevance, asker, and answerer. Additionally, rich information in the unlabeled test cQA pairs are utilized via transductive learning to enhance the representation ability of the common space. Extensive experiments on real-world datasets have well-validated the proposed model.

1 Introduction

As compared to the traditional factoid QA systems, community-based QA (cQA) systems are user-centered that leverage crowdsourcing platforms to encourage users to ask and/or answer questions, and further provide rich knowledge to search engines. Nevertheless, the quality of cQA contents varies largely due to users' diverse expertise and the low-cost posting behaviors. Low-quality cQA pairs can cause some serious problems including, but not limited to, the followings: 1) They may mislead information seekers and hence impact their search experience. 2) A large portion of storage and computing resources are wasted on low-quality contents. Instead, desired information are not acquired. 3) They decrease the stickiness of community users. Once these sites are overwhelmed by low-quality information, users will be gradually unwilling to keep browsing them. To gain insights into cQA quality, we randomly selected 100 questions and their corresponding answers from a cQA site StackExchange, and invited three volunteers to rate these cQA pairs on the scale from 1 (poor) to 5 (excellent). According to our statistics, we found that more than 45% of cQA pairs are rated as unsatisfactory (bellow 3 on average). The study result is consistent with what has been reported by Jeon et al. [2006]. To handle such issues, we target at assessing the quality of cQA pairs.

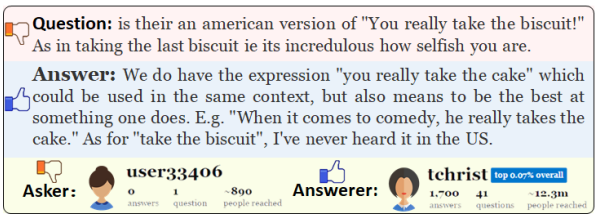


Figure 1: Aspects influence the quality of cQA pairs.

In essence, great efforts have been dedicated to predicting the quality of cQA contents [Shah and Pomerantz, 2010; Dalip *et al.*, 2013; Nie *et al.*, 2017]. Despite their success, most existing methods simply concatenated various features into single vectors first and then fed them into regression models. However, this may be suboptimal, due to: 1) **Week Capacity in Multi-view Fusion.** The quality of cQA pairs is influenced by distinct views, and each of them describes a specific but incomplete angle. We carefully went through aforementioned cQA pairs and found that their quality are determined by five main factors, including question, answer, QA relevance, asker, and answerer. Figure 1 illustrates a selected example. Even the answerer is experienced and he/she resolves the question with a well-organized answer, the cQA pair can hardly be considered as a high-quality one, as the asker is of less experience and the question is full of typos. So, overlooking the agreement among these views may cause an unfair judgement. 2) **Data Deficiency.** In most previous work, only labeled cQA pairs are utilized to train the regression model. Yet, it is expensive to gather sufficient labeled cQA pairs in practice. Instead, unlabeled cQA pairs are much cheaper to obtain, and they also contains rich information. However, these unlabeled cQA pairs are not fully explored.

Toward this end, we propose a **T**ransductive **M**ulti-**v**iew **L**earning model, dubbed as **TMvL**, to assess the cQA pair

*Heyan Huang is the corresponding author.

quality. Specifically, multi-view learning is conducted to solve the multi-view fusion problem. It seamlessly takes the view agreement into account by learning a common space shared by all views, and the desired optimal common space maintains the original intrinsic characteristics of cQA pairs in the original spaces. Meanwhile, transductive learning is applied to involve unlabeled cQA pairs into the common space learning, so that the data deficiency problem could be solved.

The main contributions of this paper are three-fold. 1) As far as we know, this is the first work applying multi-view learning to cQA pair quality assessment, where the agreement among different views of cQA pairs are fully explored. 2) We study the utility of unlabeled cQA pairs in quality assessment and propose a transductive model to fully incorporate unlabeled data. 3) Comparative experiments with both automatic and manual evaluation on our self-collected dataset well-validate the proposed model. Additionally, we have released our codes and data to facilitate follow-on researchers[1].

2 Related Work

The most related work is on answer quality prediction via maximum entropy [Jeon *et al.*, 2006]. Most follow-on efforts explored different features, such as answer content [Surdeanu *et al.*, 2008], domain knowledge [Nie *et al.*, 2015b], and embedding features [Zhang *et al.*, 2017]. Dalip et al. [2013] summarized commonly used features and analyzed their effectiveness. Beyond exploring features, other novel methods were also developed, including propagation method [Li *et al.*, 2015] and click model [Wei *et al.*, 2015]. Different from the previous efforts on exploring answer quality or question quality, we mainly focus on assessing the quality of cQA pairs.

Multi-view learning originates to learn from multiple sources or different feature subsets by considering the agreement of different views. They can be roughly categorized into three groups: co-training, multiple kernel learning, and subspace learning. Co-training alternately maximizes the mutual agreement on two distinct views. Co-EM [Nigam and Ghani, 2000], Bayesian Co-Training [Yu *et al.*, 2011], and Co-Regression [Zhou and Li, 2005] belong to this category. Multiple kernel learning [Sonnenburg *et al.*, 2005] exploits different kernels to different views, and then combines them together to enhance the performance. Subspace learning works by finding a latent subspace shared by each view, assuming that each input view is generated from the common subspace. MSNL [Song *et al.*, 2015a] , SM^{2}L [Song *et al.*, 2015b], aM^{2}L [Nie *et al.*, 2015a], and MvDA [Kan *et al.*, 2016] fall into this category.

Transductive learning is quite different from the traditional inductive learning. It leverages rich information in the testing set while training the model. Various efforts have been dedicated for unlabeled data utilization, such as Transductive SVM [Joachims, 2001], CTA [Blum and Mitchell, 2000], and SSR [Kim *et al.*, 2009]. Promising performance of transductive learning has demonstrated its effectiveness of incorporating unlabeled data.

TMvL belongs to both multi-view learning and transductive learning. Different from previous efforts, it learns a latent subspace by minimizing the disagreement between the common space and each view by jointly considering both labeled and unlabeled data.

[1]http://datapublication.wixsite.com/tmvl.

3 cQA Pair Quality Assessment

3.1 Notations

We first declare some notations used in this paper. Suppose there are N labeled and M unlabeled cQA pairs in the dataset. Each pair is described by V views. Let $\mathbf{x}_i^{(v)} \in \mathbb{R}^{D^{(v)}}$ denotes the feature vector of the i-th cQA pair in the v-th view, whereby $D^{(v)}$ refers to the feature dimension. The v-th view is represented as $\mathbf{X}^{(v)} = \left[\mathbf{x}_1^{(v)}, \mathbf{x}_2^{(v)}, \ldots, \mathbf{x}_{N+M}^{(v)}\right]^T \in \mathbb{R}^{(N+M)\times D^{(v)}}$. Quality labels of the N cQA pairs are vectorized as $\mathbf{y}^{(l)} = \left[y_1^{(l)}, y_2^{(l)}, \ldots, y_N^{(l)}\right] \in \mathbb{R}^N$, where $y_i^{(l)}$ is the quality score for the i-th cQA pair. As to the M unlabeled cQA pairs, we aim to assess their quality $\mathbf{y} = [y_{N+1}, y_{N+2}, \ldots, y_{N+M}] \in \mathbb{R}^M$, where y_i denotes the inferred quality score of the i-th unlabeled cQA pair. Our research objective is to propagate the quality scores of labeled cQA pairs to unlabeled ones. This is a transductive learning model, in which both labeled and unlabeled cQA pairs are used for training. In particular, information of both labeled and unlabeled cQA pairs are simultaneously leveraged to learn a common space shared by multiple views. The common space is denoted as $\mathbf{X}^{(0)} = \left[\mathbf{x}_1^{(0)}, \mathbf{x}_2^{(0)}, \ldots, \mathbf{x}_{N+M}^{(0)}\right]^T \in \mathbb{R}^{(N+M)\times D^{(0)}}$, where $\mathbf{x}_i^{(0)} \in \mathbb{R}^{D^{(0)}}$ indicates the representation of the i-th cQA pair in the common space.

3.2 Proposed Model

As reported in [Song *et al.*, 2015a], the squared loss usually yields good performance compared with other complex ones in the regression problem. We followed this scheme by utilizing the squared loss function to measure the training error of the labeled cQA pairs. In **TMvL**, the quality of cQA pairs is assessed relied on the representation in the common space. We thus have the empirical loss,

$$\Gamma_1 = \frac{1}{2N}\sum_{i=1}^{N}\left(y_i^{(l)} - \mathbf{w}^T\mathbf{x}_i^{(0)}\right)^2, \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^{D^{(0)}}$ is our desired coefficient vector, mapping the cQA pair representation in the common feature space to the quality value space.

The common space is induced from V different views, and each view conveys a specific aspect of cQA pairs. To informatively and comprehensively characterize each QA pair, the original intrinsic characteristics of different views are reserved in the common space. The effectiveness of Laplacian matrix has been proven in data representation and clustering [Feng *et al.*, 2017], which captures the inherent relatedness among data points, following the traditional spectral graph embedding theory. We thus turn to penalize the disagreement of normalized Laplacian matrix between the common space and each view, to ensure that the quality characteristics of different views can be preserved as much as possible.

It is formulated as,

$$\Gamma_2 = \frac{1}{2V}\sum_{v=1}^{V}\left\|\mathbf{L}^{(0)} - \mathbf{L}^{(v)}\right\|_F^2, \tag{2}$$

where $\mathbf{L}^{(k)}$ denotes the normalized Laplacian matrix, and it is calculated as,

$$\mathbf{L}^{(k)} = \mathbf{I} - \mathbf{D}^{(k)^{-\frac{1}{2}}}\mathbf{S}^{(k)}\mathbf{D}^{(k)^{-\frac{1}{2}}}, \tag{3}$$

where $\mathbf{I} \in \mathbb{R}^{(N+M)\times(N+M)}$ is an identity matrix; $\mathbf{S}^{(k)}$ and $\mathbf{D}^{(k)} \in \mathbb{R}^{(N+M)\times(N+M)}$ denote the affinity matrix and the degree matrix, respectively. In this paper, we adopt cosine similarity to measure $\mathbf{S}^{(k)}$. To decrease the complexity of the formula, we normalize $\mathbf{x}_n^{(k)}$ via enforcing $\left\|\mathbf{x}_i^{(k)}\right\|_F^2 = 1$. In this way, the (m,n)-th element of $\mathbf{S}^{(k)}$ is calculated as $S_{m,n}^{(k)} = \mathbf{x}_m^{(k)^T}\mathbf{x}_n^{(k)}$. Inspired by [Liu *et al.*, 2013], we set $S_{n,n}^{(k)} = 0$ to eliminate the self-loop problem. The degree matrix $\mathbf{D}^{(k)}$ is a diagonal matrix, and the (m,m)-th element $D_{m,m}^{(k)}$ is calculated as,

$$D_{m,m}^{(k)} = \sum_{n=1}^{N+M} S_{m,n}^{(k)}. \tag{4}$$

It is notable that the common space learning follows the main idea of transductive learning. Particularly, in the Laplacian matrix, each data point is represented by its relations with the others, including both labeled and unlabeled ones. Therefore, both labeled and unlabeled cQA pairs are equally utilized. This is why we claim that the common space $\mathbf{X}^{(0)}$ is capable of harvesting unlabeled cQA pairs to enforce its representation ability.

To incorporate the learned common space into the quality assessment task and strengthen the prediction performance, we co-regularize the empirical loss together with Eqn.(2). We then reach the final objective function $O(\mathbf{w}, \mathbf{X}^{(0)})$,

$$\begin{aligned}&\frac{1}{2N}\sum_{i=1}^{N}\left(y_i^{(l)} - \mathbf{w}^T\mathbf{x}_i^{(0)}\right)^2 \\ &+ \frac{\mu}{2V}\sum_{v=1}^{V}\left\|\mathbf{L}^{(0)} - \mathbf{L}^{(v)}\right\|_F^2 + \frac{\lambda}{2}\left\|\mathbf{w}\right\|_F^2, \\ s.t.\ &\left\|\mathbf{x}_i^{(0)}\right\|_F^2 = 1,\ i = 1, 2, \ldots, N+M,\end{aligned} \tag{5}$$

where μ and λ are both non-negative regularization parameters. Specifically, μ penalizes the disagreement between the common space and each view, and λ controls the complexity.

3.3 Optimization

We adopt the alternating strategy to minimize our objective function. We first fix $\mathbf{w}$ and optimize O w.r.t $\mathbf{X}^{(0)}$ with gradient descent. It is worth noting that the matrix $\mathbf{X}^{(0)}$ is normalized after updating each parameter to ensure $\left\|\mathbf{x}_i^{(0)}\right\|_F^2 = 1$.

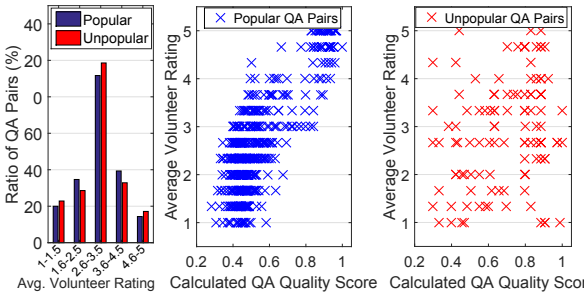


Figure 2: Results of volunteer ratings.

As to $\mathbf{w}$, we first fix $\mathbf{X}^{(0)}$, set $\frac{\partial O(\mathbf{w},\mathbf{X}^{(0)})}{\mathbf{w}}$ to zero, and we then have the closed-form solution,

$$\mathbf{w} = \left(\frac{1}{N}\mathbf{X}^{(0)^T}\mathbf{X}^{(0)} + \lambda\mathbf{I}\right)^{-1}\left(\frac{1}{N}\mathbf{X}^{(0)^T}\mathbf{y}^{(l)}\right), \tag{6}$$

where $\mathbf{X}^{(0)} = \left[\mathbf{x}_1^{(0)}, \mathbf{x}_2^{(0)}, \ldots, \mathbf{x}_N^{(0)}\right]^T$ is the representation of labeled samples in the common space, and $\mathbf{I}$ is an $N \times N$ identity matrix.

4 Experiments

4.1 Dataset Construction

We crawled data from two compact subsites of StackExchange, i.e., "English" and "Game". We collected $26,752$ questions, their corresponding $92,397$ answers, and $28,271$ users (askers and answerers) from "English" subsite. Meanwhile, we gathered $28,023$ questions, their corresponding $59,423$ answers, and $24,079$ users from "Game" subsite.

We utilized user vote information on StackExchange to construct quality labels. For a given cQA pair, the quality label considers both its answer quality s_a and question quality s_q. We employed the normalized answer vote to represent s_a, which is calculated as $s_a = \frac{v_a}{max_a(\{v_a^q\})}$, where v_a is the number of votes received by the answer a, and $max_a(\{v_a^q\})$ indicates the largest number of votes among the answers under the question q. We next utilized the number of votes v_q received by question q to infer s_q. The quality of question is normalized as $s_q = \frac{v_q}{max_q(\{v_q\})}$. We then used the average score of s_a and s_q to represent the quality of a cQA pair.

To justify the rationality of the auto-generated labels, we randomly selected 50 popular questions from "English" site with more than 5,000 viewers and 50 unpopular questions with less than 500 views. Based on these criteria, 456 and 111 cQA pairs were involved, respectively. Three volunteers were invited to annotate these cQA pairs on the scales from 1 (poor) to 5 (excellent) according to pre-determined guidelines. We calculated the inter-rater agreement among volunteers, and the pair-wise Kappa coefficients are all larger than 0.6. This demonstrates the significant alignment among them. The average score is hence regarded as the rating of the cQA pair. We observed a similar rating distribution between popular and unpopular questions, as shown in Figure 2(a), which reflects that the question popularity has less influence on the real quality of cQA pairs. Figure 2(b) and Figure 2(c) plot relations between volunteer ratings and auto-generated quality labels on popular and unpopular questions, respectively.

Dataset	Auto-Evaluation			Manual Evaluation		
	# Labeled cQA Pairs	# Unlabeled cQA Pairs		# Labeled cQA Pairs	# Unlabeled cQA Pairs	
		Training Only	Training & Testing		Training Only	Training & Testing
English	3,764	0	940	4,704	900	100
Game	2,435	0	608	3,043	900	100

Table 1: Dataset description w.r.t auto-evaluation and manual evaluation.

We found that the quality of popular cQA pairs can be better simulated, where the Pearson's correlation is 0.81, but it does not perform well on unpopular questions, where the Pearson's correlation is only 0.26. We hence selected the top 5% questions with the largest number of views to automatically construct quality labels. In total, we obtained $4,704$ and $3,043$ labeled cQA pairs from these two datasets, respectively.

4.2 Evaluation Metrics

In order to comprehensively justify our **TMvL** model, we evaluated it in two different ways, i.e., **Auto-Evaluation** and **Manual Evaluation**.

In auto-evaluation, we utilized the automatically generated labels to evaluate the performance. We randomly selected 20% cQA pairs as unlabeled samples as well as testing samples. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) [Cao *et al.*, 2017] were used as metrics.

Since auto-evaluation is fit for popular questions only, we hence manually rated a number of cQA pairs, including both popular and unpopular questions. In manual evaluation, the aforementioned $4,704$ and $3,043$ automatically labeled cQA pairs were all treated as labeled data, and we randomly selected $1,000$ cQA pairs from each subsite as unlabeled ones. It is worth noting that there is no overlap between the labeled and unlabeled cQA pairs. From these unlabeled cQA pairs, we further randomly selected 100 cQA pairs and invited three volunteers to annotate their quality scores from 1 (poor) to 5 (excellent). The pair-wise Kappa coefficients are all larger than 0.6, so the average scores of the three volunteer ratings were used as the ground truth. Since volunteer ratings are discrete values between 1 and 5, while the predicted scores are continuous between 0 and 1, Pearson's Correlation (Corr.) [Shah and Pomerantz, 2010] are utilized.

Detailed description of the data used in these two evaluation methods are displayed in Table **??**.

4.3 Feature Extraction

We intend to comprehensively assess the quality of cQA pairs from five distinct views: question, answer, cQA relevance, asker, and answerer. To accomplish this, we extracted a rich set of quality-oriented features from each view. It should be noted that the purpose of this study is not to explore the discrimination of features, so all features we used in this paper have been demonstrated effective in prior work [Li *et al.*, 2012; Surdeanu *et al.*, 2008; Dalip *et al.*, 2013].

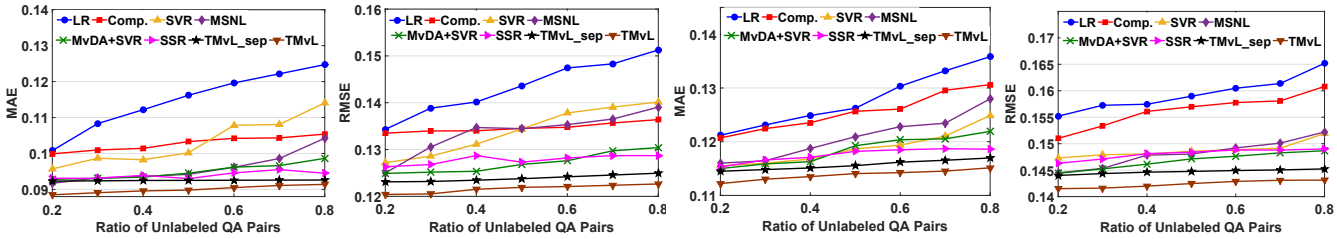
We next describe features in each view. 1) **Question View.** We extracted three kinds of features to characterize the question quality, i.e., User Generated Content (UGC) features, linguistic features, and embedding features. The number of tags, answers, user favourites, and the question age belong to UGC features. Linguistic features include the number of words, non-stop words, and sentences, the average and the maximum length sentences, the percentage of punctuation, pronouns, conjunctions, prepositions, spaces, as well as sentences starting with pronouns, articles, conjunctions, or prepositions. We also extracted a 50-dimensional sentence embedding with Para2Vec [Le and Mikolov, 2014]. 2) **Answer View.** Similar to question view, we extracted these three kinds of features. The UGC features contain the number of comments and editors. As to the linguistic features, all that described in the question view were also applicable. A 50-dimensional embedding feature was also extracted. 3) **QA Relevance View.** We extracted BM25, cosine similarity, the number of common words and common sequences. These features describe the QA relevance. Besides, the number of new adjectives, new nouns, and new verbs in answers, as well as the ratio of answer length to question length are extracted. They verify the linguistic difference. The time span between question and answer posting time are also extracted. 4) **Asker View and Answerer View.** We treated askers and answerers as two different views, but they share the same features space. We extracted features based on user achievements, such as the number of badges, reputations, and flags. Users' behaviour preference signals their attitude. We hence extracted the number of questions asked, answered, edited, and votes casted.

4.4 Baselines

We compared **TMvL** with the following baselines: 1) **LR** [Shah and Pomerantz, 2010]. This is a single-view method. We concatenated all features extracted from the five views, and then directly fed them into a logistic regression model. 2) **Comp.** [Liu *et al.*, 2011]. We trained five logistic regression models on five views separately, and fed these regression results into a final regressor to predict the quality score. This is a single-view method, as view agreement is ignored. 3) **SVR** [Drucker *et al.*, 1996]. Support Vector Regression is another single-view method, which concatenates features from different views into a single feature vector. We chose the learning formulation with the kernel of radial-basis function, as it outperforms other kernels in our experiments. 4) **MSNL** [Song *et al.*, 2015a]. Multiple Social Network Learning is a multi-view learning method, which takes both view agreement and view confidence into account. 5) **MvDA+SVR** [Kan *et al.*, 2016]. Multi-view Discriminant Analysis is another multi-view learning method. It aims to learn a single unified discriminant common space from multiple views by jointly optimizing multiple view-specific transfor-ms. We then fed feature representations of five distinct views in the common space into a **SVR** model to make further prediction. 6) **SSR** [Kim *et al.*, 2009]. Semi-Supervised Regression is a transductive learning method. It uses the information of unlabeled data to construct normal coordinates around each unlabeled point, and the normal coordinates are then employed to estimate the Hessian regularizer.

Methods	English Dataset				Game Dataset			
	MAE (Auto)	RMSE (Auto)	Corr. (Manual)	p-value	MAE (Auto)	RMSE (Auto)	Corr. (Manual)	p-value
LR	0.1008	0.1343	0.5882	5e-7	0.1212	0.1552	0.6879	2e-7
Comp.	0.0999	0.1335	0.6010	7e-7	0.1207	0.1510	0.7423	1e-7
SVR	0.0957	0.1272	0.6218	1e-6	0.1155	0.1473	0.7419	9e-6
MSNL	0.0917	0.1250	0.7116	3e-4	0.1160	0.1445	0.7700	2e-3
MvDA+SVR	0.0928	0.1249	0.7227	7e-7	0.1150	0.1444	0.7900	6e-4
SSR	0.0931	0.1263	0.6944	4e-6	0.1153	0.1463	0.7504	5e-4
TMvL_sep	0.0923	0.1231	0.7668	3e-4	0.1145	0.1440	0.8081	3e-4
TMvL	**0.0885**	**0.1204**	**0.7843**	- -	**0.1122**	**0.1415**	**0.8263**	- -

Table 2: Performance comparison between our model and baselines w.r.t 20% unlabeled data.



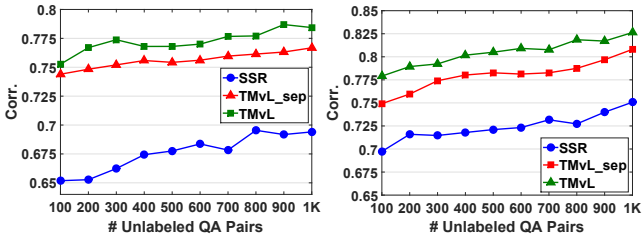
(a) MAE on English Dataset. (b) RMSE on English Dataset. (c) MAE on Game Dataset. (d) RMSE on Game Dataset.

Figure 3: Results of auto-evaluation by varying the portion of unlabeled cQA pairs on "English" and "Game" datasets, respectively.

7) **TMvL_sep**. This baseline is the variant of our proposed model by optimizing Γ_1 and Γ_2 separately. It first learns representations of cQA pairs with Γ_2. Then the representations of labeled ones are then fed into Γ_1 to train the regression model. This method belongs to transductive learning.

4.5 Overall Performance Comparison

The overall performance comparison is summarized in Table 2. We have the following observations: 1) As compared to single-view methods, i.e., **LR**, **Comp.**, and **SVR**, multi-view learning methods, i.e., **MSNL**, **MvDA**, **TMvL_sep**, and **TMvL**, perform better on both MAE and RMSE. Because the latter explores the relatedness among views to strengthen their combination. 2) Transductive learning methods, i.e., **SSR**, **TMvL_sep**, and **TMvL** surpass the inductive ones, i.e., **LR**, **Comp.**, and **SVR**. The main reason is that transductive learning-based approaches are capable of harvesting information from the unlabeled data. 3) Regarding the comparison among multi-view learning methods, **TMvL** and **TMVL_sep** outperform others. This is because that unlabeled cQA pairs are utilized to learn the common space by **TMvL** and **TMvL_sep**. This indicates that the use of unlabeled data is helpful to learn a reliable common space. 4) When comparing three transductive learning methods, we observed that **TMvL** and **TMVL_sep** perform better. Even though both labeled and unlabeled cQA pairs were used, **TMvL** and **TMVL_sep** separately considered different views and modeled the view agreement. This justifies that **TMvL** and **TMVL_sep** can well preserve the characteristics of different views. 5) Compared to **TMvL_sep**, **TMvL** co-regularizes both Γ_1 and Γ_2, so the common space learned by **TMvL** is guided by the quality label. The outstanding performance of **TMvL** demonstrates that under the supervision of labeled cQA pairs, the common space can better assess the quality of cQA pairs. 6) No matter in auto-evaluation or manual evaluation, we got quite similar results, which means that for unpopular questions, our auto-generated quality labels are reliable enough to guide the training procedure. 7) Results of pair-wise significance tests on RMSE are much smaller than 0.05, which indicates that the performance improvement is significant.



(a) English Dataset (b) Game Dataset

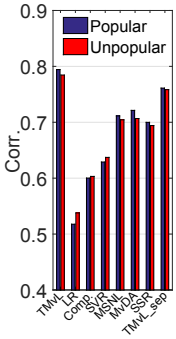
Figure 4: Transductive learning analysis with manual evaluation.

4.6 Transductive Learning Analysis

To justify the impact of incorporating unlabeled data by a transductive model, we conducted experiments by varying the number of unlabeled cQA pairs.

In auto-evaluation, we gradually increased the portion of unlabeled cQA pairs from 20% to 80%; the portion of labeled pairs therefore decreased accordingly. As shown in Figure 3, the performance of all methods gradually drops. Moreover, the performance of transductive learning methods, i.e., **SSR**, **TMvL_sep** and **TMvL** remain relatively stable as compared to others. This shows that transductive models can better alleviate the data deficiency problem.

We then performed manual evaluation to further analyze the influence of unlabeled data. We kept the labeled cQA pairs and gradually increased unlabeled ones from 100 to 1,000. Only transductive learning methods were compared here because the change of unlabeled data has no effect on inductive learning methods. The result is shown in Figure 4. We have the following observations: 1) The performance of all methods is enhanced by increasing the number of unlabeled cQA pairs. This demonstrates that transductive learning methods are able to learn a more reliable common space from unlabeled data. 2) Among these three methods, **TMvL** consistently performs the best. This is because view agreement are considered in the common space by **TMvL**.



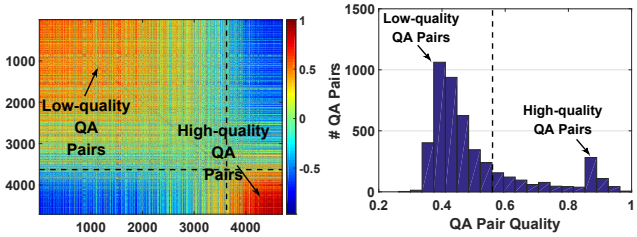
cQA Pair Content	Scores (Rating/TMvL/LR/Comp./SVR /MSNL/MvDA/SSR/TMvL_sep)
Question: What is the meaning of "ant-industry". The full sentence is "All genuine historical work is philosophy, unless it is mere ant-industry." **Answer**: You could possibly take ant-industry to mean ant-work. ... Ant work referring to something done en masse collectively if that wasn't clear. ...	2.33 / 0.42 / 0.64 / 0.52 / 0.57 / 0.52 / 0.62 / 0.54 / 0.49
Question: Does this stylistic guideline have a name? Avoid a long modifying clause that creates distance between your subject and verb. E.g., "One of his uncles who had made millions of dollars ... left John a giant inheritance." should be rewritten as "One of his uncles had made millions of dollars .... This man left John a giant inheritance." **Answer**: The guidelines are more given to disagreements of personal,... often vague in definition,... often contradictory, ... and harder to detect .... These features in themselves make them less likely to have name, and less likely again to have names that are universal. ...	4.33 / 0.80 / 0.72 / 0.68 / 0.73 / 0.66 / 0.77 / 0.63 / 0.74

(a) Evaluation (b) Two selected cQA pair examples and their corresponding quality scores calculated with different methods.

Figure 5: Performance comparison on unpopular questions with manual evaluation and some selected examples.

Methods	English Dataset			Game Dataset		
	MAE	RMSE	Corr.	MAE	RMSE	Corr.
Answer	0.1136	0.1484	0.5089	0.1467	0.1821	0.4793
+QA	0.1090	0.1430	0.5490	0.1319	0.1660	0.6069
+Answerer	0.1002	0.1363	0.6683	0.1248	0.1574	0.6759
+Question	0.0927	0.1259	0.7573	0.1156	0.1461	0.7887
+Asker	0.0885	0.1204	0.7843	0.1122	0.1415	0.8263

Table 3: Incremental multi-view integration.



(a) Similarity visualization. (b) Quality label distribution.

Figure 6: Visualization of the similarity matrix.

4.7 Evaluation on Unpopular Questions

Our quality labels were constructed with popular questions, we hence manually evaluate the performance on both popular and unpopular questions to demonstrate its applicability. The cQA pairs from those 50 popular and 50 unpopular questions annotated in section 4.1 are used as unlabeled samples, and 4,704 auto-labeled pairs are labeled ones. As the result shown in Figure 5(a), we found the followings: 1) As we mentioned in section 4.1, the Pearson's Correlation between our generated quality labels and volunteer ratings on unpopular questions is 0.26. And we found that all these comparative methods achieve higher performance. In addition, the performance on labeled and unlabeled questions is quite similar. This demonstrates that the constructed quality labels with popular questions are reasonable to supervise the model training. 2) **TMvL** stably outperformed other baselines in both two settings. It demonstrates that the proposed model is applicable to both popular and unpopular questions.

Figure 5(b) illustrates two examples and the quality assessment results of volunteers and all methods. The quality of the first one is bad, and volunteer rating is 2.33 on average, much lower than 3. Similarly, **TMvL** assesses it a low score of 0.42. While other methods tend to consider it as a high-quality one. The second example is of high quality. Even though all methods gave it high scores, the score rated by **TMvL** is much closer to human judgements.

4.8 Incremental Analysis of View Impact

To demonstrate the effectiveness of these five views in assessing cQA pair quality, we incrementally fed them into **TMvL**. Results are displayed in Table 3. Both auto-evaluation and manual evaluation were conducted, and they led to the same conclusion. It is observed that more views lead to better performance, and the best performance is obtained when all these five views are integrated. This reveals that every view contains useful information and can be effectively encoded in the common space by **TMvL**.

4.9 Common Space Visualization

To demonstrate the rationale of the common space learned by **TMvL**, we visualized similarities among cQA pairs in the common space. All cQA pairs were first sorted according to their label values, so that cQA pairs of high-quality and low-quality can be separated. We then calculated the similarity matrix, where each entry denotes the cosine similarity of two given cQA pairs in the common space $\mathbf{X}^{(0)}$. As the result shown in Figure 6(a), almost all cQA pairs are closer to the ones with similar quality. They can hence be roughly grouped into two clusters, i.e., high-quality ones and low-quality ones. After analyzing the distribution of label values, we found that these cQA pairs can also be divided into two categories, as shown in Figure 6(b). These two separations statically match, and the dot lines in these two sub-figures illustrate the same separation. This demonstrates that with the supervision of labels, the common space is more reliable and separable to assess the cQA pair quality.

5 Conclusion

This paper presents a novel transductive multi-view learning model to assess cQA pair quality. It learns an optimal common space by jointly leveraging both labeled and unlabeled training samples of the given cQA pairs with multi-facets. The common space enables each view to maintain the intrinsic properties. Comparative experiments with both automatic and manual evaluation on real-world datasets have demonstrated the promising performance of our model.

Acknowledgments

The work was mainly supported by the National High Technology Research and Development Program of China (863 Program, No. 2015AA015404) and the National Natural

Science Foundation of China (61751201). This research is part of NExT++ project, supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@Singapore Funding Initiative.

References

[Blum and Mitchell, 2000] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT'00*, pages 92–100, 2000.

[Cao *et al.*, 2017] Da Cao, Liqiang Nie, Xiangnan He, Xiaochi Wei, Shunzhi Zhu, and Tat-Seng Chua. Embedding factorization models for jointly recommending items and user generated lists. In *SIGIR'17*, pages 585–594, 2017.

[Dalip *et al.*, 2013] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pavel Calado. Exploiting user feedback to learn to rank answers in q&a forums: a case study with stack overflow. In *SIGIR'13*, pages 543–552, 2013.

[Drucker *et al.*, 1996] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alexander J. Smola, and Vladimir Vapnik. Support vector regression machines. In *NIPS'96*, pages 155–161, 1996.

[Feng *et al.*, 2017] Fuli Feng, Liqiang Nie, Xiang Wang, Richang Hong, and Tat-Seng Chua. Computational social indicators: a case study of chinese university ranking. In *SIGIR'17*, pages 455–464, 2017.

[Jeon *et al.*, 2006] Jiwoon Jeon, W Bruce Croft, Joon Ho Lee, and Soyeon Park. A framework to predict the quality of answers with non-textual features. In *SIGIR'06*, pages 228–235, 2006.

[Joachims, 2001] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *ICML'01*, pages 200–209, 2001.

[Kan *et al.*, 2016] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. *TPAMI*, 38(1):188–194, 2016.

[Kim *et al.*, 2009] Kwang In Kim, Florian Steinke, and Matthias Hein. Semi-supervised regression using hessian energy with an application to semi-supervised dimensionality reduction. In *NIPS'09*, pages 979–987, 2009.

[Le and Mikolov, 2014] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML'14*, pages 1188–1196, 2014.

[Li *et al.*, 2012] Baichuan Li, Tan Jin, Michael R Lyu, Irwin King, and Barley Mak. Analyzing and predicting question quality in community question answering services. In *WWW'12*, pages 775–782, 2012.

[Li *et al.*, 2015] Xin Li, Yiqun Liu, Min Zhang, Shaoping Ma, Xuan Zhu, and Jiashen Sun. Detecting promotion campaigns in community question answering. In *IJCAI'15*, pages 2348–2354, 2015.

[Liu *et al.*, 2011] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Evgeniy Gabrilovich, Yoelle Maarek, Dan Pelleg, and Idan Szpektor. Predicting web searcher satisfaction with existing community-based answers. In *SIGIR'11*, pages 415–424, 2011.

[Liu *et al.*, 2013] Hairong Liu, Longin Jan Latecki, and Shuicheng Yan. Fast detection of dense subgraphs with iterative shrinking and expansion. *TPAMI*, 35(9):2131–2142, 2013.

[Nie *et al.*, 2015a] Liqiang Nie, Luming Zhang, Yi Yang, Meng Wang, Richang Hong, and Tat-Seng Chua. Beyond doctors: Future health prediction from multimedia and multimodal observations. In *MM'15*, pages 591–600, 2015.

[Nie *et al.*, 2015b] Liqiang Nie, Yi-Liang Zhao, Mohammad Akbari, Jialie Shen, and Tat-Seng Chua. Bridging the vocabulary gap between health seekers and healthcare knowledge. *TKDE*, 27(2):396–409, 2015.

[Nie *et al.*, 2017] Liqiang Nie, Xiaochi Wei, Dongxiang Zhang, Xiang Wang, Zhipeng Gao, and Yi Yang. Data-driven answer selection in community QA systems. *TKDE*, 29(6):1186–1198, 2017.

[Nigam and Ghani, 2000] Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-training. In *CIKM'00*, pages 86–93, 2000.

[Shah and Pomerantz, 2010] Chirag Shah and Jefferey Pomerantz. Evaluating and predicting answer quality in community qa. In *SIGIR'10*, pages 411–418, 2010.

[Song *et al.*, 2015a] Xuemeng Song, Liqiang Nie, Luming Zhang, Mohammad Akbari, and Tat-Seng Chua. Multiple social network learning and its application in volunteerism tendency prediction. In *SIGIR'15*, pages 213–222, 2015.

[Song *et al.*, 2015b] Xuemeng Song, Liqiang Nie, Luming Zhang, Maofu Liu, and Tat-Seng Chua. Interest inference via structure-constrained multi-source multi-task learning. In *IJCAI'15*, pages 2371–2377, 2015.

[Sonnenburg *et al.*, 2005] Sören Sonnenburg, Gunnar Rätsch, and Christin Schäfer. A general and efficient multiple kernel learning algorithm. In *NIPS'05*, pages 1275–1282, 2005.

[Surdeanu *et al.*, 2008] Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Learning to rank answers on large online qa collections. In *ACL'08*, pages 719–727, 2008.

[Wei *et al.*, 2015] Xiaochi Wei, Heyan Huang, Chin-Yew Lin, Xin Xin, Xianling Mao, and Shangguang Wang. Reranking voting-based answers by discarding user behavior biases. In *IJCAI'15*, pages 2380–2386, 2015.

[Yu *et al.*, 2011] Shipeng Yu, Balaji Krishnapuram, Rómer Rosales, and R Bharat Rao. Bayesian co-training. *JMLR*, 12(9):2649–2680, 2011.

[Zhang *et al.*, 2017] Xiaodong Zhang, Sujian Li, Lei Sha, and Houfeng Wang. Attentive interactive neural networks for answer selection in community question answering. In *AAAI'17*, pages 3525–3531, 2017.

[Zhou and Li, 2005] Zhi-Hua Zhou and Ming Li. Semi-supervised regression with co-training. In *IJCAI'05*, pages 908–913, 2005.