ACRyLIQ: Leveraging DBpedia for Adaptive Crowdsourcing in Linked Data Quality Assessment

Umair ul Hassan[1], Amrapali Zaveri[2], Edgard Marx[3], Edward Curry[1], and Jens Lehmann[4,5]

[1] Insight Centre of Data Analytics, National University of Ireland, Galway, Ireland
{umair.ulhassan,edward.curry}@insight-centre.org
[2] Stanford Center for Biomedical Informatics Research, Stanford University, USA.
amrapali@stanford.edu
[3] AKSW Group, University of Leipzig, Germany
emarx@informatik.uni-leipzig.de
[4] Computer Science Institute, University of Bonn
jens.lehmann@cs.uni-bonn.de
[5] Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)
jens.lehmann@iais.fraunhofer.de

Abstract. Crowdsourcing has emerged as a powerful paradigm for quality assessment and improvement of Linked Data. A major challenge of employing crowdsourcing, for quality assessment in Linked Data, is the cold-start problem: how to estimate the reliability of crowd workers and assign the most reliable workers to tasks? We address this challenge by proposing a novel approach for generating test questions from DBpedia based on the topics associated with quality assessment tasks. These test questions are used to estimate the reliability of the new workers. Subsequently, the tasks are dynamically assigned to reliable workers to help improve the accuracy of collected responses. Our proposed approach, ACRyLIQ, is evaluated using workers hired from Amazon Mechanical Turk, on two real-world Linked Data datasets. We validate the proposed approach in terms of accuracy and compare it against the baseline approach of reliability estimate using gold-standard tasks. The results demonstrate that our proposed approach achieves high accuracy without using gold-standard tasks.

1 Introduction

In recent years, the *Linked Data* paradigm [7] has emerged as a simple mechanism for employing the Web for data and knowledge integration. It allows the publication and exchange of information in an interoperable way. This is confirmed by the growth of Linked Data on the Web, where currently more than 10,000 datasets are provided in the Resource Description Format (RDF)[6]. This vast amount of valuable interlinked information gives rise to several use cases to discover meaningful relationships. However, in all these efforts, one crippling

[6] http://lodstats.aksw.org

problem is the underlying data quality. Inaccurate, inconsistent or incomplete data strongly affects the consumption of data as it leads to unreliable conclusions. Additionally, assessing the quality of these datasets and making the information explicit to the publisher and/or consumer is a major challenge.

To address the challenge of *Linked Data Quality Assessment* (LDQA), crowdsourcing has emerged as a powerful mechanism that uses the "wisdom of the crowds" [9]. An example of a crowdsourcing experiment is the creation of LDQA tasks, then submitting them to a crowdsourcing platform (e.g. Amazon Mechanical Turk), and paying for each task that the workers perform [21, 16, 6]. Crowdsourcing has been utilized in solving several problems that require human judgment include LDQA. Existing research has focused on using crowdsourcing for detecting quality issues [21], entity linking [6], or ontology alignments [16, 14]. A major challenge of employing crowdsourcing for LDQA is to accurate responses for tasks while considering *reliability* of workers [15, 8, 5, 3]. Therefore, it is desirable to find the most reliable workers for the tasks.

In this paper, we study the problem of *adaptive task assignment* in crowdsourcing specifically for quality assessment in Linked Data. In order to make appropriate assignments of tasks to workers, crowdsourcing systems currently rely on the estimated reliability of workers based on their performance on previous tasks [15, 10]. some approaches rely on expectation-maximization style approaches to jointly estimate task responses and reliability of workers after collecting data from large number of workers [15, 11]. However, it is a difficult problem to estimate the reliability of new workers. In fact, existing crowdsourcing systems have been shown to exhibit a *long-tail* phenomena where the majority of workers have performed very few tasks [10]. The uncertainty of worker reliability leads to low *accuracy* of the aggregated tasks responses. This is called as the *cold-start problem* and is particularly challenging for LDQA, since the tasks may require domain knowledge from workers (e.g. knowledge of a language for detecting incorrectly labeled language tags).

Existing literature on crowdsourcing that addresses the cold-start problem is not applicable to LDQA due to several reasons [2, 8]. Firstly, the manual creation of *gold-standard tasks* (GSTs) with known correct responses is expensive and difficult to scale [15]. Secondly, the effects of domain specific knowledge on the reliability of workers is not considered in existing literature [15]. Moreover, assignments using social network profiles of workers require significant information about workers and their friends, which poses a privacy problem [2].

We introduce the *Adapative Crowdsourcing for Linked Data Quality Assessment* (ACRyLIQ), a novel approach that addresses the cold-start problem by exploiting a generalized knowledge base. ACRyLIQ estimates the reliability of a worker using test questions generated from the knowledge base. Subsequently, the estimated reliability is used for adaptive task assignment to the best workers. Indeed, with the generality of the DBpedia [12] (the Linked Data version of the Wikipedia), it is not difficult to find facts related to most topics or domains. As a consequence, a large quantity of domain specific and mostly correct facts can be obtained to test the knowledge of workers. Thus, the fundamental research question addressed in the paper is: *How can we estimate the reliability*

of crowd workers using facts from DBpedia to achieve high accuracy of LDQA tasks though adaptive task assignment? The core contributions of this paper are:

- A novel approach, ACRyLIQ, to generate test questions from DBpedia. The test questions are used for estimating the reliability of workers while considering the domain-specific topics associated with LDQA tasks.
- A comparative study of the proposed approach against baseline approaches on LDQA tasks using two real datasets. The first dataset considers language verification tasks for five different languages. The second dataset considers entity matching tasks for five topics: (i) Books, (ii) Nature, (iii) Anatomy, (iv) Places, and (v) Economics.
- Evaluation of the proposed and baseline approaches by employing workers from Amazon Mechanical Turk. The results demonstrate that our proposed approach achieves high accuracy without the need for gold-standard tasks.

2 Preliminaries

In this section, we introduce the core concepts used throughout this paper. Furthermore, we highlight the key assumptions associated with those concepts.

Definition 1 (Topics). *Given a Linked Data dataset, let S be a set of topics associated with the dataset. Each topic $s \in S$ specifies an area of knowledge that is differentiated from other topics.*

For instance, consider a dataset consisting of review articles about books. An article might refer to various topics such as "Books", "Political-biographies", "1960-novels", etc. We assume that similar topics are grouped together and there is minimum overlap between the topics (or topic groups) in the set S.

Definition 2 (Tasks). *Let $T = \{t_1, t_2, ..., t_n\}$ be the set of LDQA tasks for the dataset. Each task $t_i \in T$ is a multiple-choice question with an unknown correct response $r_i{}^*$ that must be generated through crowdsourcing.*

For instance, a task might ask workers to judge whether two review articles are referring to the same book. We assume that the set of tasks T is partitioned according to topics associated with the dataset; hence, each task is associated with a topic. In practice, it is possible that a task might be associated with more than one topic. In such a case, the primary topic of each task can be chosen using a relevance ranking. If there is no obvious ranking, then the primary topic of a task can be chosen arbitrarily.

Definition 3 (Workers). *Let $W = \{w_1, w_2, ..., w_m\}$ be the set of workers that are willing to perform tasks. Workers arrive in an online manner and request tasks; in addition, each worker $w_j \in W$ has a latent reliability $p_{i,j}$ on task $t_i \in T$.*

If worker w_j performs task t_i and provides the response $r_{i,j}$, then the probability of $r_i = r_{i,j}$ is $p_{i,j}$ and the probability of $r_i \neq r_{i,j}$ is $1 - p_{i,j}$. Without the loss of generality we assume that $p_{i,j} \in [0, 1]$. For instance, a worker who is

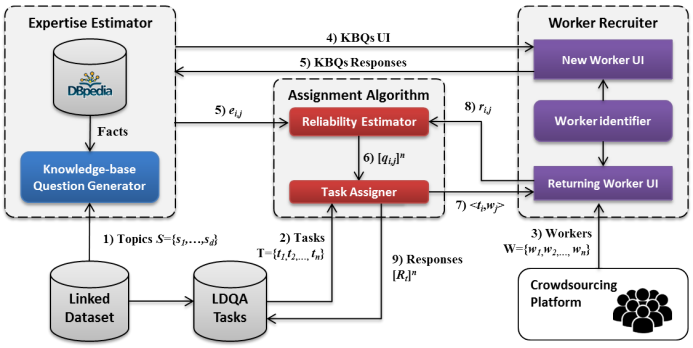

Fig. 1. An overview of the proposed ACRyLIQ approach.

well versed on the topic "1960-novels" has higher likelihood of providing correct responses to tasks associated with the same topic.

Let $R_i = \{r_{i,j} \mid w_j \in W_i\}$ be the set of responses collected from the workers $W_i \subset W$ assigned to the task t_i. We use majority-vote on R_i to generate the response $\hat{r}_i$. The goal of an assignment algorithm is to find the best workers for a task such that the estimated response $\hat{r}_i$ is accurate. Therefore, the assignment algorithm must estimate the reliabilities of workers. The set of workers W_i for a task t_i can be chosen such that $\sum_{w_j \in W_i} p_{i,j}$ is maximized. In case of arbitrary reliabilities, the assignment algorithm may not be able to find good worker. We assume that the reliability of a worker on tasks associated with the same topic remain approximately the same. That is $p_{i,j} \sim p_{i',j}$ for all w_j when t_i and $t_{i'}$ are associated with the same topic.

Definition 4 (Gold-standard tasks). *The subset of tasks $T_G \subset T$ with known correct responses (most often from the same dataset).*

Existing approaches for adaptive task assignment use gold-standard tasks to estimate reliabilities of workers which imposes additional costs for the collection of correct responses from domain experts [8, 3]. Furthermore, it is difficult to generate gold-standard tasks for complex or knowledge-intensive tasks [15].

Definition 5 (Knowledge Base). *A set of facts F related to the set of topics S where each fact belongs to exactly one topic.*

We assume access to a knowledge base that contains facts F related to the dataset of LDQA tasks. Similar to the partitioning of tasks, the facts in the knowledge base are divided in to $|S|$ partitions. Next, we describe a novel approach for estimating the reliabilities of workers by exploiting a knowledge base.

3 Reliability Estimation using Knowledge Base

Figure 1 illustrates the proposed approach, for reliability estimation and adaptive task assignment, that uses DBpedia as a generalized knowledge base of

facts[7]. First, the topics S are used for selecting facts from DBpedia and generating test questions, referred to as *knowledge base questions* (KBQs). Then, each new worker w_j is given the KBQs and worker's responses to KBQs are used to generated estimated reliabilities $q_{i,j}$. Finally, the estimated reliabilities are used for assigning LDQA tasks to workers and estimation of task responses. Note that a similar approach can be applied to knowledge bases other than DBpedia. The number of facts in a knowledge base can be in millions which raises the KBQ selection challenge: *How to choose a set of facts from the knowledge base such that the reliability of a worker on the KBQs highly correlates with their reliability on LDQA tasks*? The goal is to minimize the difference between the estimated reliability $q_{i,j}$ and the true reliability $p_{i,j}$.

3.1 KBQs Selection Problem

Recent research has shown that the reliability of a worker tends to be comparable on similar topics [3]. Therefore, we propose to use the similarity between tasks and facts to address the KBQs selection problem. The intuition is that the similarity quantifies the influence of workers' response to both facts and tasks. Since the similarity can be defined in terms of textual comparisons or detailed semantics, we detail the similarity measure used in this paper in Section 5. Given the similarity measure $sim(t, f)$, next we formalize the KBQs selection problem.

Let Q be the set of KBQs generated from facts F in the knowledge base. The number of KBQs is fixed at Φ to control the overhead costs of reliability estimation. Based on their associated facts, the KBQs in Q are also divided into $|S|$ partitions. Let Q_s be the set of KBQs selected for topic s. The probability that a KBQ is selected for topic s is $\mathbb{P}(Q_s) = |Q_s|/\Phi$. We define the *entropy* of the set Q according to a measure based on Shannon entropy [17], that is $\mathbb{H}(Q) = -\sum_{s \in S} P(Q_s) \cdot \ln P(Q_s)$. The intuition is to generate a diverse set of KBQs. A higher value of entropy means that more topics are covered with equal number of KBQs; hence, it is desirable to maximize the entropy of Q.. Besides entropy, the objective is to generate KBQs that have high influence on the tasks. Influence is the positive correlation between the accuracy of worker responses to KBQ and the accuracy of the same worker on tasks. The next section details a parametric algorithm that addresses the KBQs selection problem.

3.2 KBQs Selection Algorithm

We devise a greedy algorithm for the KBQs selection problem, as shown in Algorithm 1. The algorithm assumes availability of a similarity measure $sim(t, f)$ between LDQA tasks and facts in the knowledge base. The algorithm starts with an empty set of facts Q and then iteratively selects Φ facts from the knowledge base F. For each new fact f_k, the algorithm calculates the difference between the entropy of Q and the entropy of $Q \cup f_k$ (Line 5). Then it selects the fact $\hat{f}$ that maximizes the entropy difference and the similarity with the tasks, and β is used

[7] A DBpedia triple is considered a fact.

Algorithm 1 KBQ Selection Algorithm

Require: T, S, F, Φ, β
1: $Q \leftarrow \emptyset$
2: **for** $i = 1, ..., \Phi$ **do**
3: $F \leftarrow F - Q$
4: **for** $f_k \in F$ **do**
5: $\Delta_k = \mathbb{H}(Q \cup f_k) - \mathbb{H}(Q)$
6: **end for**
7: $\hat{f} = \text{argmax}_{f_k \in F} \beta \cdot \Delta_k + (1 - \beta) \sum_{t_i \in T} sim(t_i, f_k)$
8: $Q \leftarrow Q \cup \hat{f}$
9: **end for**
10: **return** Q

as the similarity-entropy trade-off parameter (Line 7). The computational complexity of Algorithm 1 is $\mathcal{O}(\Phi|F||T|)$. Understandably, a performance bottleneck can be the number of facts in the knowledge base. This necessitates an effective pruning strategy to exclude facts that are very different from tasks or have little benefit for reliability estimation. Section 5.2 discusses one such strategy that is employed to reduce the search space when selecting facts from DBpedia.

4 Adaptive Task Assignment

Given the set of KBQs, we extend an existing adaptive task assignment algorithm that uses gold-standard tasks to estimate worker reliabilities [3]. This algorithm also serves as the baseline during the evaluation of our proposed approach. Algorithm 2 lists our algorithm for adaptive task assignment that uses KBQs for reliability estimates. The algorithm expects a set of tasks T, a set of KBQs Q, and three control parameters (i.e. λ, α, γ). The parameter λ specifies the number of unique workers $|W_i|$ to be assigned to each task. The similarity-accuracy trade-off parameter is α and the number of iterations is γ. The algorithm consists of two distinct phases: (i) offline initialization and (ii) online task assignment.

The offline initialization phase (Lines 1-18) consists of following steps. The algorithm starts by combining LDQA tasks and KBQs (Lines 2-3) and calculates their similarity based on the topics shared between them (Line 4). For instance, LDQA tasks or KBQs belonging to the same topic are assigned a similarity value between 0 and 1. Next, the algorithm normalizes the similarity scores (Lines 5-9). The similarity scores in matrix $\hat{\boldsymbol{Z}}$ are further weighted using the parameter α to control the effect of similarity on reliability estimation (Lines 10-17). Parameter γ controls the number of iterations used for the adjustment of similarity scores.

The online task assignment phase (Lines 19-31) proceeds in iterations as workers arrive dynamically and request tasks. The task assignment process stops when all tasks have received λ responses. Each dynamically arriving worker w_j requests C_j tasks (Line 22). If the worker w_j is requesting tasks for the first time, then the set of KBQs is assigned to the worker (Line 24). Based on the responses to the KBQs, an estimated reliability vector $\boldsymbol{p_j}$ is generated for the worker w_j

Algorithm 2 Adaptive Assignment Algorithm

Require: $T, Q, \lambda, \alpha, \gamma$

1: $T \leftarrow T \cup Q$ {Combine tasks and KBQs}
2: $n \leftarrow |T|$
3: $\boldsymbol{Z} \leftarrow TopicSimilarityMatrix(T)$ {Topic similarity matrix}
4: $\boldsymbol{D} \leftarrow [0]^{n \times n}$
5: **for** $i = 1, ..., n$ **do**
6: $\quad D_{i,i} = \sum_{j=1}^{n} Z_{i,j}$
7: **end for**
8: $\hat{\boldsymbol{Z}} \leftarrow \boldsymbol{D}^{-1/2} \boldsymbol{Z} \boldsymbol{D}^{-1/2}$
9: **for** $t_i \in T$ **do**
10: $\quad \boldsymbol{p_i} \leftarrow [0]^n$
11: $\quad p_{i,i} \leftarrow 1$
12: $\quad \boldsymbol{q_i} \leftarrow \boldsymbol{p_i}$
13: $\quad$ **for** $g = 1, ..., \gamma$ **do**
14: $\quad\quad \boldsymbol{p_i} \leftarrow \frac{1}{1+\alpha} \boldsymbol{p_i} \hat{\boldsymbol{Z}} + \frac{\alpha}{1+\alpha} \boldsymbol{q_i}$
15: $\quad$ **end for**
16: **end for**
17: $c \leftarrow 0$ {Initialize assignments counter}
18: $R \leftarrow \emptyset$ {Initialize response set}
19: **for** $c < n\lambda$ **do**
20: $\quad (w_j, C_j) \leftarrow getNextWorker()$ {Worker requests C_j tasks}
21: $\quad$ **if** w_j is a new worker **then**
22: $\quad\quad$ Assign Q KBQs to worker
23: $\quad\quad \boldsymbol{q_j} \leftarrow ObservedAccuracy(Q)$
24: $\quad\quad \boldsymbol{p_j} \leftarrow \sum_{q_{i,j}} q_{i,j} \cdot \boldsymbol{p_i}$
25: $\quad$ **end if**
26: $\quad \mathcal{T} = \{\tau | \tau \subset T, |\tau| = C_j\}$
27: $\quad T_j^* = \text{argmax}_{\tau \in \mathcal{T}} \sum_{t_i \in \tau} p_{i,j}$
28: $\quad$ Assign T_j^* to worker w_j to get R_j responses.
29: $\quad c \leftarrow c + C_j$
30: $\quad R \leftarrow R \cup R_j$
31: **end for**
32: **return** R

(Lines 25-26). The set of tasks, for which the w_j has highest reliabilities, is assigned to the worker (Lines 28-30). At the end of an iteration, the assignment counter and response sets are updated. The computational complexity of the offline initialization phase is $\mathcal{O}(n^2)$ and the online assignment phase is $\mathcal{O}(n)$.

5 Evaluation Methodology

For the purpose evaluation, we collected responses to KBQs and LDQA from real workers on Amazon Mechanical Turk. Since repeated deployment of the assignment algorithm with actual workers is known to be difficult and expensive [3, 18], we employed a simulation-based evaluation methodology to compare the

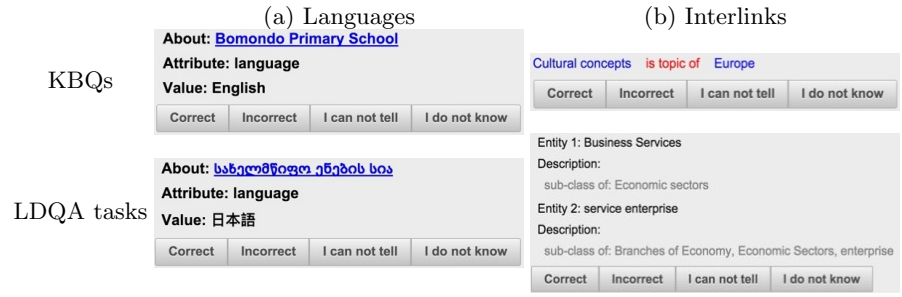

Fig. 2. Examples of KBQs and LDQA tasks for the Languages and Interlinks datasets.

algorithms using collected responses. Each run of the algorithm is initialized with specific tasks and worker conditions.

5.1 LDQA Tasks

The LDQA tasks were based on two real-world datasets, as shown in Figure 2. The datasets are summarized below:

- **Languages dataset:** These tasks represent the syntactic validity of datasets [20], that is, the use of correct datatypes (in this case correct language tags) for literals. The tasks are based on the LinkedSpending[8] dataset, which is a Linked Data version of the OpenSpending project[9]. The dataset contains financial budget and spending data from governments from all over the world. As OpenSpending does not contain language tags for its entities, the 981 values of LinkedSpending entities only contain plain literals. In an effort to accurately identify the missing language tags, we first applied an automated language detection library[10] to generate a dataset containing the entities and the corresponding language. Out of the 40 distinct languages detected, 25 entity language pairs were randomly chosen to generate tasks. The correct responses for tasks were created with the help of language translation tools.
- **Interlinks dataset:** These tasks represent the interlinking quality of a dataset [20], specifically about the presence of correct interlinks between datasets. The tasks were generated from the Ontology Alignment Evaluation Initiative[11] (OAEI) datasets that cover five topics: (i) Books, (ii) Geography, (iii) Nature, (iv) Economics and (v) Anatomy. Each task required workers to examine two entities along with their corresponding information and evaluate whether they are related to each other (e.g. if they are the same). The correct responses for these tasks are available along with the OAEI datasets.

[8] `http://linkedspending.aksw.org/`
[9] `https://openspending.org/`
[10] `https://github.com/optimaize/language-detector`
[11] `http://oaei.ontologymatching.org/`

Table 1. Summary of experiment parameters and their default values (in bold font).

Parameter	Description	Values
λ	Assignment size per task	**3**, 5, 7
α	The similarity-accuracy trade-off parameter	0, **0.5**, 1
β	The similarity-entropy trade-off parameter	0.25, **0.5**, 0.75
γ	The number of iterations	0.25, **0.5**, 0.75
n	Number of LDQA tasks i.e. $\|T\|$	**15**
m	Number of crowd workers i.e. $\|W\|$	**60**
Φ	The number for KBQs or GSTs per worker	5, **10**, 15

5.2 KBQs Selection with Pre-pruning

Since DBpedia contains more than three billion facts[12], it was essential to devise a pruning strategy to assist the KBQs selection process. In order to reduce the search space, we employed pre-pruning with the help of a string similarity tool. We used the LIMES tool [13], which employs time-efficient approaches for large-scale link discovery based on the characteristics of metric spaces. In particular, we used LIMES to find resources from the DBpedia dataset similar to the entities mentioned in the tasks. The triples associated with similar resources were used as facts for KBQ generation. As a pruning strategy, we could restrict the resources for each domain by specifying the particular class from DBpedia. For example, the resources for the topic "Books" were restricted to instances of the DBpedia class "Book". For the Languages dataset, the specification of the language in the LIMES configuration file assisted in selecting resources that were in that particular language. LIMES also supports the specification of using a specific string similarity metric, as well as the corresponding thresholds. In our case we used the "Jaro" similarity metric [19] and retrieved all the resources above threshold of 0.5. Figure 2 shows the examples of knowledge base questions generated for both datasets, as summarized below:

- **Language dataset:** A set of 10 KBQs for the Languages dataset was generated from DBpedia. Each test question asks the worker to identify whether a value has a correct language tag.
- **Interlinks dataset:** The KBQs were focused towards estimating the expertise of a worker on the topics associated with the Interlinks dataset. Triples with DBpedia property "is subject of" were used to generate the 10 KBQs. Each question required workers to identify whether one entity is related to (i.e. is subject of) another entity.

5.3 Compared Approaches & Metrics

We evaluated the performance of three reliability estimation approaches: (i) the proposed KBQ approach, (ii) the existing GST approach and the baseline *randomly generated estimates* (RND) approach. The following metrics were used to report the performance of algorithms:

[12] As of 2014 `http://wiki.dbpedia.org/about`

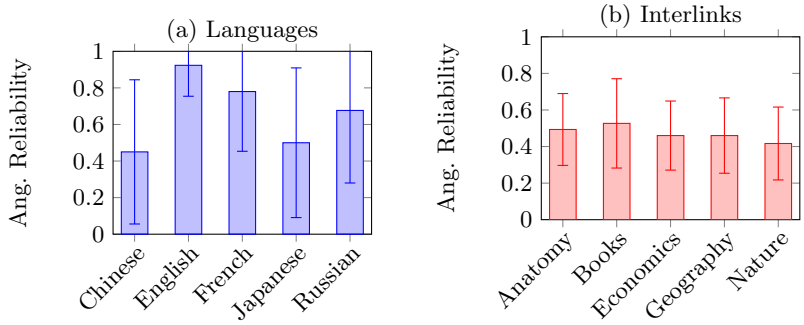

Fig. 3. Average reliability of workers on all 25 tasks for both datasets.

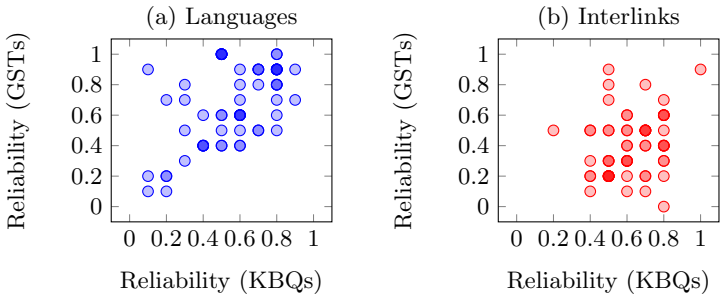

Fig. 4. Reliability of the 60 real workers on KBQs and GSTs.

- *Average Accuracy*: The primary metric for the performance is the average accuracy of the final aggregated responses over all tasks.
- *Overhead Costs*: The total overhead costs paid to workers due to the KBQs or GSTs used for estimating the worker reliabilities.

6 Experimental Results

Revisiting our research question, we aim to estimate the reliability of a worker using KBQs in an effort to assign the best workers for LDQA tasks. In the following, we first report the results of data collected from real workers and then present the results of simulation experiments performed for the evaluation of the proposed approach. Table 1 shows the experimental parameters and their default values.

6.1 Diverse Reliability of Crowd Workers

The KBQs and LDQA tasks were posted on the a dedicate web server[13]. We used Amazon Mechanical Turk to hire 60 Master workers. The workers were

[13] `http://dataevaluation.aksw.org`

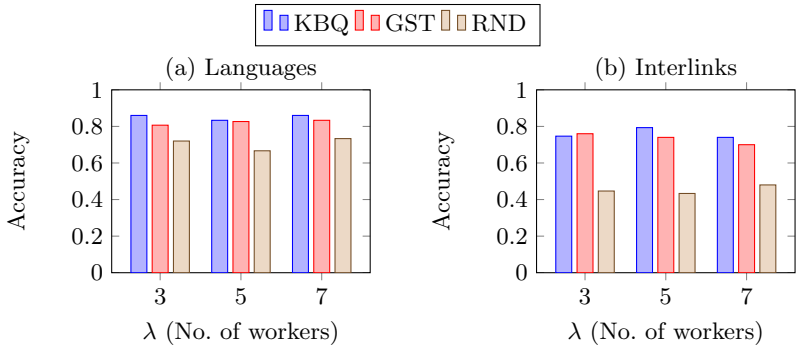

Fig. 5. Accuracy of the reliability estimation approaches for Languages and Interlinks.

paid a wage at the rate of $1.5 for 30 minutes spent on the tasks. Worker was first asked to provide background information such as: the region that they belong to, their self-assessed knowledge about the five topics (of the Interlinking dataset), the amount of years they have spoken each language (of the Languages dataset) and their native language. Then the worker was asked to answer the 10 KBQs for each dataset. The sequence of KBQs was randomized for each worker. Finally, the worker was asked to respond to the set of 25 LDQA tasks for each dataset. Workers took nine minutes, on the average, to complete the background information, the KBQs, and the LDQA tasks.

We used the first 10 tasks in both datasets as the gold-standard tasks. Figure 3 shows the average reliability of workers in terms of the languages in the Languages dataset and topics in the Interlinks dataset. Note that the workers are less reliable on Asian languages and their standard deviation of reliability is high. Workers exhibit high reliability on European languages with low standard deviation. The average reliability is low across the topics in the Interlinks datasets. Figure 4 shows the relationship between the reliability of workers on KBQs and GSTs. The Pearson correlation between the two reliabilities is 0.545 and 0.226 for the Languages and Interlinks datasets, respectively.

6.2 Accuracy of Compared Approaches

We compared the average accuracy of the proposed approach against two baseline approaches: RNDs and GSTs. We also varied the λ parameters to study its effects on the performance of each approach. Figure 5 shows the accuracy on the Language and Interlinking tasks, based on 30 runs of each approach under the same settings. In general, the accuracy for both the KBQ and the GST approach is better than the baseline RND approach. This underlines the effectiveness of the adaptive task assignment algorithm in finding reliable workers for LDQA tasks. We compared the accuracy of the KBQ approach against the RND and GST approaches using the t-test, on the Languages dataset. The difference between the KBQ approach and the RND approach is significant with

Table 2. Effects of parameters Φ and β on the accuracy for the Interlinks dataset.

	Overhead costs budget			Similarity-entropy trade-off		
λ	Φ=5	Φ=10	Φ=15	β=0.25	β=0.5	β=0.75
3	0.709 ±0.029	0.716 ±0.033	0.718 ±0.034	0.684 ±0.028	0.780 ±0.025	0.749 ±0.032
5	0.716 ±0.033	0.758 ±0.030	0.758 ±0.028	0.760 ±0.030	0.760 ±0.027	0.740 ±0.027
7	0.744 ±0.032	0.727 ±0.026	0.742 ±0.025	0.733 ±0.025	0.733 ±0.030	0.736 ±0.034

$t(178) = 13.745$ and $p < 0.05$. The difference between the KBQ approach and the GST approach is also significant with $t(178) = 3.719$ and $p < 0.05$.

These results establish the effectiveness of adaptive task assignment in exploiting the diverse reliability of workers for the improvement of accuracy. In the case of the Languages dataset, the average accuracy of RND is closer to both KBQ and GST, although still statistically lower. This can be attributed to the lower variance of worker reliability of the Languages dataset in comparison to the Interlinks dataset.

6.3 Effects of Algorithm Parameters

We also studied the effects of the three algorithm parameters (i.e. Φ, α, β) on the average accuracy. For this purpose, we used a subset of the DBpedia resources using the pre-pruning strategies discussed earlier (c.f. Section 5.2). These resources were utilized to generate 74,993 KBQs that were used for the experiments. The similarity values between the KBQs and LDQA tasks were calculated using the Jaro-Winkler similarity measure. We simulated the answers of the workers on the 74,993 KBQs by training a logistic regression model from their answers to the 10 KBQs presented to them earlier. The model accuracy was more than 72% on test instances. We used this model to analyze the effects of different parameters on the performance the proposed algorithm.

The parameter Φ defines the budget for the overhead costs due to KBQs. Table 2 shows that the accuracy increases with increase in Φ; however, the relative increase is marginal. This indicates that even at the small cost Φ of estimating reliabilities through KBQs, the assignment algorithm achieves high accuracy. The parameter β controls the similarity-entropy trade-off. As shown inTable 2, the highest accuracy is achieved for $\beta = 0.5$. Meaning that any non-extreme value for the similarity-entropy trade-off parameter is sufficient. Similar results were observed for similarity-accuracy trade-off parameter α. In general, the conservative values of these parameters do not have significant effects on performance. However, this might change with a larger number of tasks with multiple topics.

7 Discussion & Limitations

The majority of existing literature on on adaptive task assignment in crowdsourcing considers GSTs for the cold-start problem [3, 8, 22]. Generating GSTs in itself is a difficult and expensive process [15]. Especially when the accuracy of

task responses is not measurable. A key strength of our proposed approach is the applicability to such scenarios. It provides a quick and inexpensive method of estimating the reliability and expertise of workers. This approach is particularly suited for complex or knowledge-intensive tasks.

Our approach has three main limitations. First, the assumption that both facts and tasks are partitioned according to the same set of topics. In practice, this assumption can be relaxed by using a mapping between topics of facts and topics of tasks. A similar approach was employed for alignment of topics for the Interlinks dataset. Second, the approach assumes that the majority of the facts, that are used for the generation of KBQs, are correct. If a high percentage of incorrect facts are used for generating KBQs then our approach can misjudge the reliability of workers on tasks. Third, it assumes that the domain topics are mutually exclusive. This underlines that need for reconsideration of the entropy measure when the domain topics are overlapping.

The experiments presented in this paper is also limited in terms of scalability. In the case of DBpedia, pre-pruning can be utilized to limit the facts to the core DBpedia ontology and SKOS concepts. The facts can also be filtered according to the ratings of their associated articles in Wikipedia. The evaluation is also limited in terms of the overhead costs of the KBQs selection algorithm. The Languages and Interlinks dataset represent two types of LDQA tasks which can also be seen as a limitation of the experimental evaluation. However, an extension of the proposed approach to other types of LDQA tasks should be straight forward.

8 Related Work

At a technical level, specific Linked Data management tasks have been subject to crowdsourcing, including entity linking [6], ontology alignment [16], and quality assurance [21, 1]. However, none of these proposals consider adaptive task assignment or the cold-start problem. Noy et al. performed a comparative study of crowd workers against student and domain experts on ontology engineering tasks [4, 14]. Their study highlighted the need for improved filtering methods for workers; however, they did not propose algorithms for KBQs generation or adaptive task assignment.

Within the literature on crowdsourcing, several approaches have been proposed for adaptive task assignment. Ho, Jabbari, and Vaughan proposed primal-dual techniques for adaptive task assignment of classification tasks using GSTs [8]. Their approach estimates reliability of a worker against different types of tasks instead of topics. Zhou, Chen, and Li proposed a multi-armed bandit approach for assigning top-K workers to a task, and their approach also uses GSTs [22]. Another approach focused on dynamic estimation of worker expertise based on conformity of workers with the majority responses [18]. Ipeirotis, Provost, and Wang proposed an approach for separating worker bias from reliability estimation [11]. Such an approach is complimentary to our algorithm for reducing the influence of spammers on task responses. Oleson et al. proposed a manual audit approach to quality control in crowdsourcing by generating gold-standard

tasks with different types of errors previously observed in different tasks [15]. By comparison, our approach focuses on automated selection of knowledge base questions for quality control in crowdsourcing. Hassan, O'Riain, and Curry used a hybrid approach of self-rating and gold-standard tasks for estimating the expertise of workers [5]. By comparison, our approach uses DBpedia facts for estimation of worker expertise.

9 Conclusion & Future Work

In this paper, we presented ACRyLIQ, a novel approach to estimate the reliability of crowd workers. ACRyLIQ supports the adaptive task assignment process for achieving high accuracy of Linked Data Quality Assessment tasks using crowdsourcing. The proposed approach leverages a generalized knowledge base, in this case DBpedia, to generate test questions for new workers. These test questions are used to estimate the reliability of workers on diverse tasks. ACRyLIQ employs a similarity measure to find good candidate questions, and it uses an entropy measure to maximize the diversity of the selected questions. The adaptive task assignment algorithm exploits the test questions for estimating the reliability. We evaluated the proposed approach using crowdsourced data collected from real workers on Amazon Mechanical Turk. The results suggest that ACRyLIQ is able to achieve high accuracy without using gold-standard tasks.

As part of the future work, we plan to apply our approach to larger datasets within multiple domains. We also plan to further investigate the relationship between the reliability of workers and the semantic similarity of facts and tasks. A detailed study is needed to understand the relationship between the reliability of workers and the semantic similarity of DBpedia facts and crowdsourcing tasks.

Acknowledgement. This work has been supported in part by the Science Foundation Ireland (SFI) under grant No. SFI/12/RC/2289 and the Seventh EU Framework Programme (FP7) from ICT grant agreement No. 619660 (WATERNOMICS).

References

[1] Maribel Acosta et al. "Crowdsourcing linked data quality assessment". In: *The Semantic Web–ISWC 2013*. Springer, 2013, pp. 260–276.

[2] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. "Pick-a-crowd: tell me what you like, and i'll tell you what to do". In: *Proceedings of the 22nd international conference on World Wide Web*. 2013, pp. 367–374.

[3] Ju Fan et al. "iCrowd: An Adaptive Crowdsourcing Framework". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM. 2015, pp. 1015–1030.

[4] Amir Ghazvinian, Natalya Fridman Noy, Mark A Musen, et al. "Creating mappings for ontologies in biomedicine: simple methods work." In: *AMIA*. 2009.

[5] Umair Ul Hassan, Sean O'Riain, and Edward Curry. "Effects of expertise assessment on the quality of task routing in human computation". In: *Proceedings of the 2nd International Workshop on Social Media for Crowdsourcing and Human Computation., Paris, France*. 2013.

[6] Umair Ul Hassan, Sean O'Riain, and Edward Curry. "Leveraging Matching Dependencies for Guided User Feedback in Linked Data Applications". In: *Proceedings of the 9th International Workshop on Information Integration on the Web*. ACM Press, 2012, pp. 1–6.

[7] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web Into a Global Data Space*. Vol. 1. Morgan & Claypool Publishers, 2011.

[8] Chien-Ju Ho, Shahin Jabbari, and Jennifer W Vaughan. "Adaptive task assignment for crowdsourced classification". In: *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. 2013, pp. 534–542.

[9] Jeff Howe. In: *Wired Magazine* 14.6 (June 2006).

[10] Panagiotis G Ipeirotis. "Analyzing the amazon mechanical turk marketplace". In: *XRDS: Crossroads, The ACM Magazine for Students* 17.2 (2010), pp. 16–21.

[11] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. "Quality management on amazon mechanical turk". In: *Proceedings of the ACM SIGKDD workshop on human computation*. ACM. 2010, pp. 64–67.

[12] Jens Lehmann et al. "DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia". In: *Semantic Web Journal* 6.2 (2015), pp. 167–195.

[13] Axel-Cyrille Ngonga Ngomo and Sören Auer. "LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data". In: *Proceedings of IJCAI*. 2011.

[14] Natalya F Noy et al. "Mechanical turk as an ontology engineer?: using microtasks as a component of an ontology-engineering workflow". In: *Proceedings of the 5th Annual ACM Web Science Conference*. ACM. 2013, pp. 262–271.

[15] David Oleson et al. "Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing." In: *Human computation* 11.11 (2011).

[16] Cristina Sarasua, Elena Simperl, and Natalya F Noy. "Crowdmap: Crowdsourcing ontology alignment with microtasks". In: *The Semantic Web–ISWC 2012*. Springer, 2012, pp. 525–541.

[17] Claude Elwood Shannon. "A mathematical theory of communication". In: *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1 (2001), pp. 3–55.

[18] Alexey Tarasov, Sarah Jane Delany, and Brian Mac Namee. "Dynamic estimation of worker reliability in crowdsourcing for regression tasks: Making it work". In: *Expert Systems with Applications* 41.14 (2014), pp. 6190–6210.

[19] William Winkler. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage". In: *Proceedings of the Section on Survey Research Methods (American Statistical Association)*. 1990, pp. 354–359.

[20] Amrapali Zaveri et al. "Quality Assessment for Linked Data: A Survey". In: *Semantic Web Journal* 7.1 (2016), pp. 63–93.

[21] Amrapali Zaveri et al. "User-driven quality evaluation of DBpedia". In: *Proceedings of the 9th International Conference on Semantic Systems*. ACM. 2013, pp. 97–104.

[22] Yuan Zhou, Xi Chen, and Jian Li. "Optimal PAC multiple arm identification with applications to crowdsourcing". In: *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 2014, pp. 217–225.