A Crowdsourcing Method for Obtaining Rephrased Questions

Nobuyuki Shimizu
Yahoo Japan Corporation
nobushim@yahoo-corp.jp

Atsuyuki Morishima
University of Tsukuba
mori@slis.tsukuba.ac.jp

Ryota Hayashi
University of Tsukuba
ryota.hayashi.2014b@mlab.info

Abstract

We propose a method for obtaining and ranking paraphrased questions from crowds to be used as a part of instructions in microtask-based crowdsourcing. With our method, we are able to obtain questions that differ in expression yet have the same semantics with respect to the crowdsourcing task. This is done by generating tasks that give hints and elicit instructions from workers. We conducted experiments with data used for a real set of gold standard questions submitted to a commercial crowdsourcing platform and compared the results with those from a direct-rewrite method.

Introduction

Microtask-based crowdsourcing is a promising approach to solve problems in many applications. A typical procedure for using a *crowdsourcing platform* is as follows: First, *requesters* design *microtasks* that have *questions* to be asked to *workers*. Then, they put the microtasks into a crowdsourcing platform. Examples of such platforms include Amazon Mechanical Turk and Yahoo! Crowdsourcing. Finally, workers perform the microtasks and the platform receives the results.

One of the essential problems in microtask-based crowdsourcing is how to design questions for microtasks. In this paper, we propose a method for obtaining a variety of questions that are different but have the same semantics. The semantics of a question is defined by the *gold standard data*, which is a set of expected answers to the question with associated data items. For example, the semantics of “Is this green tea?” is defined by the gold standard data that states that the answer is yes if and only if the photo associated with the question is green tea. “Is this green tea?” and “Is this Japanese tea?” are expected to have the same semantics.

Obtaining different questions with the same semantics has a wide range of interesting applications, especially in microtask-based crowdsourcing. First, when planning crowdsourcing tasks, quality control measures must be considered. In a conventional survey or questionnaire, several questions are often related to each other in a way that an answer to one question implies an answer to another (Lewis and Sauro 2009). For example, questionnaires may include a mix of positive and negatively worded statements, asking for the same thing in different ways. Such a construction is intended to reduce acquiescence bias; a type of response bias in which users tend to agree with a statement when in doubt; therefore, possibly answering “yes” to all the questions (Sauro and Lewis 2011). In designing crowdsourcing tasks, inserting differently worded questions expressing the same concept would help requesters to remove workers who do not produce thoughtful or adequate responses, forming a basis for ensuring quality of the results. Second, rephrased questions with the same semantics will be beneficial when finding questions that were used for microtasks in the past. Service providers for crowdsourcing platforms often have a database that stores a set of questions submitted to their platforms. Using a set of rephrased questions to retrieve past questions in the database will help providers find a useful set of past instances when designing questions for their customers.

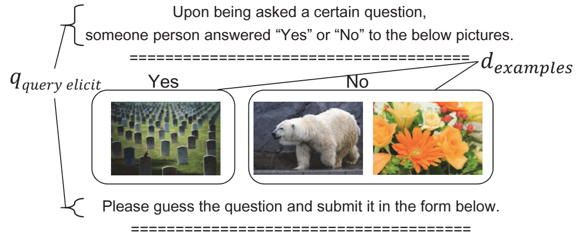
Upon being asked a certain question,
someone person answered “Yes” or “No” to the below pictures.
=====================================
Please guess the question and submit it in the form below.
=====================================

Figure 1: QE-Task

The idea behind our proposed method is that we use the gold standard data to generate *question-eliciting tasks (QE-Tasks)* to elicit questions from workers by showing them a hint and asking them to guess the questions. The contributions of this paper are as follows. To the best of our knowledge, this paper is the first to address the problem of obtaining questions for microtasks. Generating QE-Tasks is a novel approach to obtain appropriate questions. The approach is also interesting in that it is required to appropriately exploit inference capabilities of the crowd.

Formalization and the Overview

Microtasks. A microtask is a graphical interface connected to a database from which workers receive a question and data and to which they submit the inputs. In this paper, a microtask is represented as a pair (q, d_i), where q is the ques-

Original	Rephrased questions	Why better?
Is this a seasoning?	Is this something added to foods?	Positive items include Jam and similar foods/flavorings that are not usually called seasonings
Is this an instant food?	Can you begin to eat it in ten minutes after you start cooking?	Cooking some positive items require hot water or appliances other than microwave ovens. Some workers believe that instant food require such appliances for cooking.

Figure 2: Examples of better rephrased questions

tion common to a particular set of tasks, and $d_i \in D_q$ (the data set for q) is a data item associated with $t_{q,i}$. We omit q and use t_i if q is obvious from the context.

Gold standard data. Some of the tasks, called *gold standard questions*, are associated with *gold standard data*, which are answers that are given in advance by requesters. The gold standard questions are used to filter out inappropriate workers, such as spammer, and to find ones that are qualified for the tasks. In our formalization, we model gold standard data as the function $C_q : D_q \to \{\text{yes}, \text{no}\}$.

Overview of method. The proposed method takes as inputs a set D_q of data items and gold standard data C_q and outputs a list of rephrased questions $q'_1 \ldots q'_n$. We elicit questions by showing the gold standard data to crowds. We call this type of task QE-Tasks, which is a microtask that shows workers several data items in D_q and their answers defined by C_q and asks them to guess a question that, together with each d_i of the shown data items provided, produces the same answer as $C_q(d_i)$. Figure 1 is an example. A possible result elicited from a worker in this case could be "Is the object in the picture man-made?"

Query-Eliciting Tasks

A QE-task asks workers to guess a question about the data item, with the condition that its answer must also agree with the gold standard data. Figure 1 shows an example. The task is denoted by $(q_{\text{query elicit}}, d_{\text{examples}})$, where each component is as follows:

- $q_{\text{query elicit}}$ = "Upon being asked a certain question, some person answered yes to the left group of data items and no to the right group of data items. Please guess a question and submit it in the form below."
- d_{examples} contains data items in D_q grouped by $C_q(d_i)$ (i.e., yes and no).

In general, given C_q and D_q, we generate a set of QE-Tasks with Y positive and N negative examples taken from the gold standard data. To construct $QE\text{-}Tasks_q(Y, N)$, we need to choose the examples in some way (e.g., random sampling). In our experience, a very small number of samples (such as one to three sample) worked well.

For simplicity, this paper deals with only the cases in which possible answers to the original question are yes or no, but we can easily extend the proposed method to deal with more general cases.

Preliminary Results

We compared the results from our proposed method to those from the direct-rewrite method that asks workers to directly rewrite the original questions. While our method outputs a more diverse set of rephrased questions, the direct-rewrite method outputs many rephrased questions that faithfully preserve the semantics of the original question. This is because the proposed method asks workers to infer questions solely based on examples. The output diversity of our proposed method depends on the original question.

Our method obtained rephrased questions that is similar to the original question "Is this tea?," while it obtained those that were not similar to the original questions "Is this a seasoning?" and "Is this an instant food?" One of the possible reasons is that the latter two are not so clear concepts for workers even if dictionaries give their definitions. Interestingly, the method often obtained better questions when original ones were not good. Figure 2 shows examples and explains why they are better. The point is that workers for the proposed method give question sentences based on the gold standard data, and not constrained by the original question sentence.

Given these results, we argue that the proposed method is effective when the original gold standard question is incorrect or vague for workers, while the direct-rewrite method is effective when the original question is guaranteed to be reasonably good. Rephrased questions obtained with our method tend to have differing expressions compared to directly asking workers to rewrite the questions.

Acknowledgments

We are grateful to the project members of Yahoo! Crowdsourcing, including Masashi Nakagawa and Manabu Yamamoto. This work was supported by JSPS KAKENHI Grant Number 24250012 in part.

References

James R. Lewis and Jeff Sauro. The factor structure of the system usability scale. In *Proceedings of the 1st International Conference on Human Centered Design: Held As Part of HCI International 2009*, HCD 09, pages 94–103, Berlin, Heidelberg, 2009. Springer-Verlag.

Jeff Sauro and James R. Lewis. When designing usability questionnaires, does it hurt to be positive? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2215–2224, New York, NY, USA, 2011. ACM.