

Worker Viewpoints: Valuable Feedback for Microtask Designers in Crowdsourcing

Ryota Hayashi
University of Tsukuba
1-2, kasuga
Tsukuba, Ibaraki, Japan
ryota.hayashi.2014b@mlab.info

Nobuyuki Shimizu
Yahoo Japan Corporation
Kioi Tower 1-3,
Chiyoda-ku, Tokyo, Japan
nobushim@yahoo-
corp.jp

Atsuyuki Morishima
University of Tsukuba
1-2, kasuga
Tsukuba, Ibaraki, Japan
mori@slis.tsukuba.ac.jp

ABSTRACT

One of the problems a requester faces when crowdsourcing a microtask is that, due to the underspecific or ambiguous task description, workers may misinterpret the microtask at hand. We call a set of such interpretations *worker viewpoints*. In this paper, we argue that assisting requesters to gather a worker's interpretation of the microtask can help in providing useful feedback to designers, who may restate the task description if necessary. In our method, we create a corpus of viewpoints annotated with the types of viewpoints that reflect the logical structure embedded in them. Our experimental results suggest that the logic-oriented annotation is effective in choosing useful viewpoints from a possibly huge set of collected viewpoints, in the sense that removing viewpoints of particular types did not affect the quality of revised task instructions. We also show that the logic-oriented annotation can perform comparably with an entropy-based method, without several workers performing the same task in parallel.

Keywords

crowdsourcing, data quality, text analysis

1. INTRODUCTION

Microtask-based crowdsourcing is an online labor market where requesters submit small tasks and workers receive small amounts of money in return for completing them. One of the essential problems in microtask-based crowdsourcing is how to design questions for microtasks. To obtain solid, quality data from crowdsourcing, microtask must be well-designed so that the task description would not allow workers to have differing interpretations of how to process the task at hand. However, ambiguous or underspecific task description is quite common in microtask crowdsourcing, as it is very difficult for the requester to foresee all possible viewpoints that thousands of participating workers could have. This biases the task results, and, in some cases, makes the large portion of them useless. In fact, the percentage of correct answers to an ambiguous question in the experiment shown in Section 3.1 was

only 75.5%. Since there are almost always subtle ambiguities that may go undetected even by skilled requesters, providing them with feedback about differing viewpoints is essential. In order to notify a requester the existence of differing viewpoints, we developed a method that automatically selects useful viewpoints from ones provided by workers and presents them to the requester, relying on him/her to make the final decision whether the reason for the judgments are valid. Useful viewpoints let requesters understand how workers interpret the instruction and where there is ambiguity. This approach is unique in that it intends to improve task instructions, not to directly select good quality data. The approach can be combined with any other approaches to improve data quality.

Our contributions are three-fold.

Method for Collecting Worker Viewpoints. We suggest modifying a microtask to have an entry at its end to ask the workers the reason for their judgment. For example, in a microtask, a worker is asked to judge whether or not a picture contains offensive contents. Upon looking at the picture of a newborn baby, the worker judges the picture not to be offensive. He provides the answer, "No, it is not offensive." We then ask for the reason for the judgment. For example, the worker may provide his reason "Babies are cute and no one dislikes babies." In this paper, a viewpoint is defined as the reason for the judgment together with the judgment itself. In this case, a viewpoint is the pair ("No, it is not offensive", "Babies are cute and no one dislikes babies"). Giving viewpoints are optional and we paid no additional cost for the tasks used in our experiments.

Annotation Guideline for Labeling and Selecting Useful Viewpoints. Since we crowdsource giving worker viewpoints, picking up "more useful" viewpoints from them is a problem that needs to be addressed. One form of the "usefulness" is generality, i.e., how many data items the viewpoint can be applicable to. For example, consider the following viewpoints with a negative judgment: "A picture of Sunset is not usually considered offensive." and "It is just a fruit. It is neither offensive nor sexual." In these cases, collected viewpoints simply mention the name of the things that are not offensive. However, there is a very large list of things that would not be offensive. As listing all of them would be impossible, such viewpoints are useless for a requester. The following is a more useful one. "There is no nudity or violence involved in the picture. Not offensive." In this case, the reason for the judgment states the criteria that the worker used to make the judgment, instead of just naming an object that is not offensive. Thus, choosing generic viewpoints is an important criteria because they directly show how workers interpret the question.

As we will show, generality of viewpoints is strongly related to logical structure of the obtained viewpoints. Therefore, we propose an annotation guideline for labeling viewpoints with types

©2017 International World Wide Web Conference Committee (IW3C2), published under Creative Commons CC BY 4.0 License.
WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04
<http://dx.doi.org/10.1145/3041021.3051147>



representing their logical structures. We then create a corpus of viewpoints annotated with the types, which will be open to public¹. Note that the corpus is not of questions, but viewpoints. Therefore, the corpus does not depend on question sentences, but on the type of question, which, in our case, questions for classification tasks. We do not have to develop corpora for different question sentences. **Evaluations.** This paper shows a variety of evaluation results. First, we show a subjective evaluation where crowds are shown viewpoints and asked to determine if the reason stated in the viewpoint is applicable for judging another data of the same kind. The results clearly show that our scheme is useful for choosing more generic viewpoints from others.

Second, we asked people whether viewpoints were useful when revising task instructions and compared the proposed logic-based method for choosing useful viewpoints with another promising method that chooses viewpoints for microtasks whose answer distributions (i.e., entropies) are large. The results show that viewpoints are useful and the logic-based method is comparable with the entropy-based one in the quality of chosen viewpoints. This is interesting because the logic-based method analyzes viewpoint texts only, without requiring each task to be performed by many workers.

Finally, we show that we are able to use machine learning to predict the type of viewpoints. To support requesters, we propose the use of support vector machine trained with the annotated corpus with maximal substring features. The experimental results show that the learned classifier is capable of predicting with accuracy of 88.3%. While the classification accuracy has much to be desired considering the high inter-annotation agreement rate, it is a good starting point for practical use with requesters.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 defines viewpoints and its types. It then explains the procedure for collecting and annotating the data, and shows the statistics of the types as well as the inter-annotator agreement rate. Section 4 conducts a subjective evaluation and determines which types are useful. Section 5 explains the classifier and shows its classification performance. Section 6 concludes.

2. RELATED WORK

To the best of our knowledge, no prior studies have been conducted on assisting requesters solicit worker viewpoints, as well as selecting useful viewpoints for requesters in microtask crowdsourcing. In the area of crowdsourcing, the most relevant prior research was conducted by [9]. In their research, authors state that “Non-obvious attributes are not necessarily easily nameable, but nonetheless they play a systematic role in people’s interpretation of images. Clusters of related non-obvious attributes, called interpretation dimensions, emerge when people are asked to compare images, and provide important insight on aspects of social images that are considered relevant.” The authors then propose a procedure for discovering non-obvious attributes using crowdsourcing. While their motivation is similar to ours, the profound difference between [9] and our work is that, while [9] focus on images, we focus on discovering non-obvious attributes of task descriptions.

In the area of natural language processing, the closest technique that comes to our minds is that of presupposition detection. In presupposition detection, the task is to find a presupposition in a question. Presuppositions are propositions that take some information as given. For example, the question, “when did you marry”, presupposes that the respondent is married. Since someone who has

never married does not know what to respond, the answers may become random.

In order to avoid having unintentional presupposition in designing a questionnaire, presupposition detection methods have been developed [13, 11]. Presupposition detection is a part of textual inference [12] and there are various tasks where one tries to tell one sentence that can be inferred from the other sentence. While these textual inference tasks focus only on the semantics of written texts and facts associated with them, this work is very different; we focus on the requesters’ needs in order to classify the writers’ comments. The usefulness of workers’ comments to microtask designers is an abstract concept, much more so than concrete facts and relations stated in texts.

Finally, data quality is an important issue in crowdsourcing. There are different opportunities for improving data quality of task results. Before posting tasks to a crowdsourcing site, we have chances to improve the task design and to pick up workers who we expect would give good quality data [4, 2]. In this phase, our viewpoints can be used for revising original task questions when the previous task results are not expected ones. After receiving task results, we have chances to filter out inappropriate results or workers and aggregate results for improving data quality [7, 1]. Although this paper does not focus on this phase, using viewpoints in this phase is an interesting future work. For example, we can filter out task results with viewpoints that are too short.

3. VIEWPOINT ANNOTATION

In this paper, a *viewpoint* is a sentence stating why a worker gave the answer to a task question. For example, a viewpoint in answering to “Is it coffee?” is “It is black liquid in a cup. Therefore, it is coffee.”

This section presents our procedure for building a viewpoint corpus by collecting viewpoints and labeling them with the viewpoint types. We will use the corpus to automatically annotate viewpoints for other sets of microtasks.

3.1 Collecting Viewpoints

Tasks. First, we built a set of microtasks which workers perform to give their viewpoints in answering questions. For that purpose, we used a set of pictures of goods sold by ASKUL¹, a large e-commerce site in Japan, and generated microtasks each of which asks workers whether an item shown in a picture belongs to a specific category.

Because the site has a large number of goods classified in a hierarchical category, we would have too many microtasks if we used all the goods in the site. Therefore, we made microtasks for goods under the subcategories of “Drink/Foods” category, each of which (1) has at least three subcategories in it so that we have a variety of goods, and (2) does not have a category name consisting of two other category names (such as “coffee and tea”) or too generic names such as “gifts.” As a result, we obtained 28 pictures in seven categories, and the questions corresponded to the category names (Figure 1). For each question, we displayed one of the 28 pictures. Thus, we had $7 \times 28 = 196$ original microtasks.

Second, we added to each task the questions for collecting workers’ viewpoints (Figure 1). If the possible judgments are “Yes,” “No,” and “Unsure,” we added entry field for reasons for judgments “Yes” and “No.” If a worker chooses “Unsure”, he was asked to enter possible reasons for both judgments. Note that the pair of a reason and the judgment comprises a viewpoint.


¹https://crowd4u.org/ja/projects/viewpoint_corpus

¹<http://www.askul.co.jp/>

Is this coffee?
Is this a carbonated drink?
Is this Japanese tea?
Is this tea?
Is this green tea?
Is this a seasoning?
Is this an instant food?

Table 1: Questions used for building our corpus

Question: Is this coffee?



Please choose "Yes", "No" or "Unsure".

If you chose "Yes", please write the reason for "Yes".
 If you chose "No", please write the reason for "No".
 If you chose "Unsure", please write reasons for both answers assuming that the answer is correct.

Reason for "Yes":

Reason for "No":

Figure 1: Example of task asking for viewpoints

Task Submission. Each worker was given a set of five microtasks at a time and paid 2 yen (about 2 cents) for doing it. Therefore, we added four dummy microtasks to the 196 microtasks so that the total number (i.e., 200) can be divided by five to generate 40 microtask sets, where a microtask set is the unit of submission to Yahoo! Crowdsourcing².

We submitted 20 duplicates of the microtask sets to Yahoo! Crowdsourcing so that we would obtain 4,000 viewpoints at most through 800 microtask sets. Each worker was allowed to perform four sets of microtasks at most. The number of workers was 387. As a result, we obtained 1,413 viewpoints, because entering their viewpoints was not mandatory in the microtasks. Meanwhile, the percentage of correct answers to questions, "Is this an instant food?", "Is this a seasoning?", "Is this Japanese tea?" was 85.2%. In particular, the percentage for "Is this an instant food?" was only 75.5%.

3.2 Notation

In the remainder of the paper, we use the following notation for the result of the tasks for collecting viewpoints. Let W be a set of the workers, Q be a set of question sentences, $A = \{Yes, No\}$ be a set of possible judgments, and $C = A \cup \{Unsure\}$ be a set of possible choices. Then, each instance in the result can be represented as a tuple (w, q, d, a, r) where r denotes a reason for the judgment a that a worker $w \in W$ entered in response to the question q with a shown data item d . If w chooses "Unsure" for q , she enters for all judgments in A and we will have both $(w, q, d, "Yes", r)$ and $(w, q, d, "No", r)$.

A viewpoint can be thought to consist of *reasons* and a *conclusion* and we can represent it in a logical form such as $r \rightarrow c$. For

²<http://crowdsourcing.yahoo.co.jp/>

example, in the viewpoint "It is black liquid in a cup. Therefore, it is coffee," r is "it is black liquid in a cup" and c is "it is coffee."

Note that we can obtain a viewpoint $r \rightarrow c$ from each collected instance (w, q, d, a, r) where c is a positive (negative) statement for q if a is "Yes" ("No"). For example, c for the example above is derived from a question (e.g., "Is it coffee?") and a judgment (e.g., "Yes"). Therefore, we use $r \rightarrow (q, a)$ to denote a viewpoint if necessary.

3.3 Types of Viewpoints

We define the types of viewpoints in terms of this logical form. We use "N" to denote a reason or conclusion having a negative meaning and "P" to denote those having a positive meaning. From the combinations of N and P, we have four types of *simple viewpoints*. Examples are as follows.

PP ($r \rightarrow c$): It has a label having the term "Coffee". Therefore, it is coffee.

NP ($\neg r \rightarrow c$): It is coffee, because there is no evidence against it.

PN ($r \rightarrow \neg c$): It is curry. Therefore, it is not coffee.

NN ($\neg r \rightarrow \neg c$): It is not coffee, because it is not made from coffee beans.

In addition, there are two types of *composite viewpoints*.

PNP ($r_1, \neg r_2 \rightarrow c$): It has a label having the term "Coffee" and there is no evidence to argue that it is not coffee. Therefore, it is coffee.

PNN ($r_1, \neg r_2 \rightarrow \neg c$): It is not coffee because it is clear liquid but not black.

Note that in composite viewpoints, we do not assume any particular logical connector (such as And and Or) among sentences in their reasons. The only requirement is that the reason has both positive and negative statements in it. We call a viewpoint like "It is coffee because it is coffee" a *tautology*.

DEFINITION 1 (TAUTOLOGY). Given a viewpoint of type *PP* ($r \rightarrow c$) or *NN* ($\neg r \rightarrow \neg c$) in which $r = c$, we call it a *Tautology*.

The following two labels are added to indicate if a viewpoint is a tautology or not. As the first letter in PP or NN that represents the r part carries practically no content, we reduce the label to one letter.

P ($c \rightarrow c$): It's coffee. So, yes, it is coffee.

N ($\neg c \rightarrow \neg c$): No, it is not coffee because it is not coffee.

When the NN part of PNN contains tautology, we classify it to be PN.

After collecting viewpoints, we manually annotated each viewpoint with viewpoint types. Remember that a viewpoint $r \rightarrow (q, a)$ is obtained from each instance (w, q, d, a, r) we obtained from the task; When $a = "Yes"$, we annotate the viewpoint obtained by the instance with one of PP, NP, and PNP as it has a conclusion that is positive. When $a = "No"$, we annotated the viewpoint with one of PN, NN, and PNN. The exception is that we annotated a viewpoint with "Wrong" in the followings cases.

- The reason is given in the entry field of the reason for "Yes" ("No") although her answer to the microtask is "No" ("Yes"),
- The reason is not a reason (such as "None" and "Is this curry?"), or
- The reason is meaningless in the given context (For example, the question is "Is this tea?", her answer is "No" and the reason is "Yes, this is a seasoning!").

Type	ID	Question	Answer	Sentence in the entry field
<i>PP</i>	PP1	Is this Japanese tea?	Yes	It has the term "Powder Green Tea" on the package. Therefore, this is Japanese tea.
	PP2	Is this a carbonated drink?	Yes	We see the term 'Cider' on the package
<i>NP</i>	NP1	Is it an instant food?	Yes	Because we do not cook it with a traditional way such as putting leaves into a pot, I think the answer can be yes
<i>PN</i>	PN1	Is it coffee?	No	Because dressing is not coffee
	PN2	Is this tea?	No	It is coffee.
<i>NN</i>	NN1	Is this green tea?	No	Since milk tea does not contain green tea leaves
	NN2	Is this a seasoning?	No	A seasoning is something that adds a flavor to foods
<i>PNN</i>	PNN1	Is this coffee?	No	Because this is herb tea and contains no caffeine, this is not coffee
	PNN2	Is this an instant food?	No	Since an instant food is a food that is easy to cook and the beans must be roasted before eating, I think the answer is no.

Table 2: Annotation Examples

Table 2 gives examples to explain how we annotated the viewpoints.

PP2. "We see the term Cider on the package" is the reason. The answer is "Yes". Therefore, it is PP. Note that we can obtain the conclusion even if the sentence in the entry field does not have an explicit conclusion.

PN2. The answer is No and we can interpret this sentence as "Because this is coffee, this is not tea." Therefore, "This is coffee" is the reason and the type is PN.

NN2. This is a complicated case. Although "A seasoning is something that adds a flavor to foods" does not contain any negative expression, it is NN. The reason is this. First, since the question is "Is this a seasoning?" and the answer is "No", "It is not a seasoning" is the conclusion and the viewpoint is either PN or NN. Second, the sentence "A seasoning is something that adds a flavor to foods" means "if it is a seasoning, it adds flavor to foods" and its contraposition is "if it does not add any flavor to foods, it is not a seasoning". Since the answer is No, this sentence is exactly her viewpoint. Therefore, "if it does not add any flavor to foods" is the reason and it is NN. In general, if (1) the question is "Is this X", (2) the reason contains a necessary condition, and (3) the answer is "No", we conclude the viewpoint is of type NN.

PNN2. The sentence contains two reasons "An instant food is a food that is easy to cook" and "the beans must be roasted before eating", and the conclusion "I think the answer is no." The combination of the first reason and the answer has the same pattern as NN2 and is NN. The combination of the second reason and the answer is a typical PN. Therefore, it is PNN as a whole. If we have more than one reason in a sentence, we determine the types of (sub)viewpoints and combine them to determine the (composite) type of the viewpoint.

3.4 Inter-annotator Agreement and Corpus Statistics

In this section, we report the inter-annotator agreement and summary statistics for our annotated corpus. The collected viewpoints were manually annotated by two judges. We measured inter-annotator agreement and the κ value, a standard measure for showing agreements, was 0.968, showing excellent agreements between two annotators [3]. There were only 28 difficult cases where annotators disagreed in the type of viewpoints. The high agreement shows the clarity of the annotation guideline.

PP	(P)NP	PN	NN	PNN	P	N	W	Total
247	2	824	110	95	23	4	108	1,413
17.5%	0.1%	58.3%	7.8%	6.7%	1.6%	0.3%	7.6%	100%

Table 3: Numbers and percentages of viewpoints of different types

Table 3 shows the numbers and percentages of viewpoints of different types. There is a large number of viewpoints labeled as PN. W stands for wrong. NP and PNP each appears only once in the corpus. As the table shows, there are few viewpoints of types NP and PNP. This is not surprising and we can justify this. Given a viewpoint $\neg r \rightarrow c$ of type NP, its contraposition is $\neg c \rightarrow r$. In many cases, there are too many things that do not satisfy c and it is difficult to give a simple phrase to describe the common attribute of those things, that can be used for r . For example if c is "it is a coffee", it is not easy to give a simple phrase to describe the common attribute of the things that is not coffee.

4. SUBJECTIVE EVALUATION

To evaluate the effectiveness of the logic-oriented annotation scheme to select generic viewpoints, we crowdsourced judging whether each viewpoint is applicable to a variety of data items. We call this a subjective evaluation of our annotation scheme.

4.1 Crowdsourced Experiments

Tasks. Remember that in order to obtain a viewpoint $r \rightarrow (q, a)$, we asked workers why they gave the answer a to the question q with the shown data item d . The subjective evaluation tasks ask other workers to judge whether the viewpoint is useful to answer the question q with another data item $d' (\neq d)$.

To generate the subjective evaluation tasks, we used the viewpoints obtained in Section 3 and the *gold standard data* that contains a correct answer for each data item. More specifically, the tasks were generated in the following way.

Step 1. For each viewpoint $r \rightarrow (q, a)$ obtained from (w, q, d, a, r) , we randomly selected $d' (\neq d)$ whose answer was a in the gold standard data. The condition means that the answers to q with d and d' are the same. We did not generate any task for viewpoints whose types are "wrong" since viewpoints of the type do not make any sense.

Step 2. For each d' chosen in Step 1, we generated a subjective evaluation task (Figure 2). It has two questions; The first question is the same as q . We need it because we want to know whether the worker agrees on the conclusion of the gold standard data. The second question is to ask whether the shown viewpoint is applicable to d' assuming that the worker agrees on the conclusion.

We generated ten such tasks with different d' s for each viewpoint. Note that the more positive answers to the second question we obtain for a viewpoint, the more generic the viewpoint is. Conversely, the viewpoint that obtains only negative answers to the second question is too specific so that it can be applied to d only.

Workers. In order to take a majority vote, we recruited nine unique workers for each task by Yahoo! crowdsourcing platform. The number of workers who performed our tasks was 1058.

type	# viewpoints with ≥ 1 agree	# majority voted useful	percentage
PP	247	96	38.9%
NP	1	0	0.0%
PNP	1	1	100.0%
PN	824	115	14.0%
NN	110	48	43.6%
PNN	95	22	23.1%
P	23	22	95.7%
N	4	3	75.0%

Table 4: Numbers and percentages of viewpoints voted useful by the crowds

Results. Table 4 shows the result. Each line represents one viewpoint type. In each line, the second column shows the number of viewpoints (of the type) each of which has at least one worker agreed on its conclusion. The second and third columns show the numbers and the percentages of viewpoints for which the majority of workers stated that the shown viewpoint was applicable to d' .

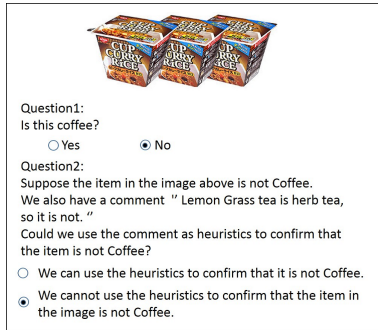


Figure 2: Subjective evaluation task

An interesting finding is that there is a clear difference in the percentage of viewpoints the workers thought were applicable to other data items. Key observations are as follows:

- The percentages of the applicable viewpoints of types P and N are clearly higher than others. The reason that viewpoints of the types are tautologies such as “This is an instant food because this is an instant food” (Section 3.3).
- The percentage for type PN is lower than those for Types PP and NN. Many of viewpoints of type PN were specific to the shown instances. For example, the viewpoint “This is not a carbonated drink, because this is tea” is applicable only when the shown data item is tea. In other words, given a viewpoint $r \rightarrow c$ of type PN, the set of data items represented by r was small so that d' was not included in the set in many cases. On the other hand, rs for types PP and NN tended to be more generic, such as “This is not food,” and were not too specific to a particular data item.
- Viewpoints of type PNN, such as “this is tea and since tea is not food, this is not an instant food,” is a combination of viewpoints of type PN and NN and thus has reasons that were both specific to and independent of an instance. Interestingly, a number of workers ignored the part specific to a particular instance, and the viewpoints showed higher percentages.

Theoretically, the percentages must be high when the data sets represented by r and c are similar to each other. Therefore, the experimental result suggests that r in viewpoints of Types PP, NN and

PNN explains the characteristics of c well, and directly represents each worker’s interpretation of c . As explained in Section 3.4, we usually have few viewpoints of Types NP and PNP, and removing them does not reduce the number of viewpoints much. On the other hand, removing viewpoints of Type PN can be an effective way to reduce the number of viewpoints.

To conclude, the logic-oriented annotations can serve as an important clue to choose viewpoints when there are too many collected viewpoints.

4.2 Verification of the Subjective Evaluation Results

We conducted an experiment to verify whether the result of the subjective evaluation suggesting that viewpoints of Type PN are not useful since they are not generic is valid. We compared the quality of task results after rewriting task questions in the following two settings.

- Setting A: show workers all collected viewpoints when they rewrite task questions.
- Setting B: show workers all collected viewpoints except those of type PN when they rewrite task questions.

Procedure. First, we asked workers to rewrite questions. We used the three questions that led to low-quality answers in Section 3.1 - “Is it an instant food?” “Is this a seasoning?” and “Is this Japanese tea?”. We asked workers to rewrite the three questions in Settings A and B. Each worker rewrote each of the three questions in different settings. For example, A worker rewrote “Is this an instant food?” in Setting A, “Is this a seasoning?” in Setting B and “Is this Japanese tea?” in Setting A. Since the number of all such combinations is six, we recruited six workers to write questions. With the six combinations of three original questions, we obtained $6 \times 3 = 18$ revised questions in total. Namely, six revised questions (three revised in Setting A and three revised in Setting B) for each original question.

Then, we combined the revised 18 questions with 28 pictures to generate $18 \times 28 = 504$ tasks, and submitted them to Yahoo!Crowdsourcing. Each worker was given a set of six microtasks at a time and paid 3 yen (about 3 cents) for doing it. Each worker was allowed to perform three sets at most.

We submitted 20 duplicates of the microtasks to Yahoo! Crowdsourcing so that we would obtain answers for $504 \times 20 = 10,080$ microtasks at most. The number of workers was 620.

Results. Table 5 shows the percentage of correct answers to the tasks with the original and revised questions. We found that in both settings, the quality of revised question results became much better than those for original questions, with the number of “unsure” being slightly increased. Importantly, removing PN-type viewpoints did not have huge impact on the quality of revised questions, although the percentage of removed viewpoints is 58.3% (see Table 3). This result clearly shows that our scheme is useful for choosing more generic viewpoints from others, and that the queries revised by people using collected viewpoints allow us to obtain high-quality task results even if the original ones were ambiguous.

4.3 Comparison with Entropy-based Method

The entropy-based approach is another promising approach that assumes viewpoints of workers on ambiguous tasks would be useful for revising questions. In this approach, we identify microtasks whose entropies are large (i.e., a variety of answers are given to the same task) then select viewpoints associated with them. We experimentally compared the proposed logic-based method with an entropy-based one. The results show that the logic-based method

Original question	Original	Setting A	Setting B
Is this an instant food?	75.5% (4.3%)	88.8% (4.5%)	90.1% (1.1%)
Is this a seasoning?	88.9% (1.4%)	96.6% (4.0%)	94.6% (1.5%)
Is this Japanese tea?	91.2% (1.7%)	94.4% (5.3%)	92.5% (2.6%)

Table 5: Percentages of correct answers to the tasks with original and revised questions. The percentages in the parentheses are those of “unsure”.

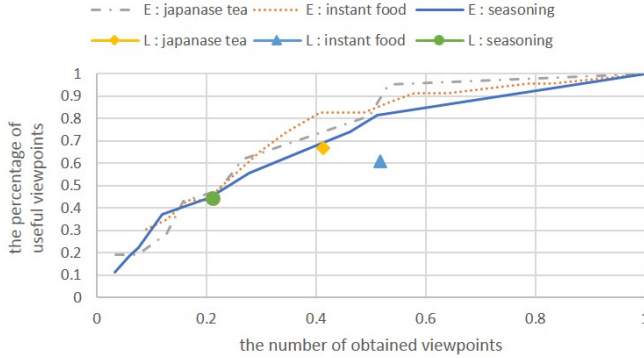


Figure 3: Results of the comparison experiment. L (E) means the Logic-based (Entropy-based) method

can perform comparably without several workers performing the same task in parallel.

Procedure. We showed six people the task instructions and the viewpoints collected in Section 3 and asked them to improve the task instructions. Remember that we submitted 20 duplicates of each microtask to obtain 20 different answers for every microtask. Therefore we can compute the entropy of obtained answers to choose useful viewpoints in the entropy-based method. The number of obtained viewpoints was 1,413. We then asked them which viewpoints were useful for improving the task instructions. They answered that 9.3 viewpoints were useful on average. Finally, we computed how many useful viewpoints were included in the results of our proposed method and the entropy-based method.

Results. Figure 3 shows the result. The x-axis is the number of shown viewpoints. In the entropy-based method, we rank the viewpoints according to the entropy of the submitted answers to the task question for which each viewpoint was given. Let t and a be a microtask and an answer to it, respectively. Given a probability function $P_t(a)$ that represents how often a appears in the answers to t , the entropy $H(t)$ is computed by $H(t) = -\sum_a P_t(a) \log P_t(a)$ [8]. The y-axis is the percentage of useful viewpoints included by the accumulated set of viewpoints. As shown in the figure the quality of the outputs of the proposed method is comparable with those of the entropy-based method.

The result is interesting and encouraging because it shows the logic-based method can choose useful viewpoints without requiring each task to be performed by more than one worker. For instant-food question, the logic-based method showed 24.7 percent inferior performance. We carefully reviewed the viewpoints and found that some PN viewpoints are not specific to particular instances and sometime useful for revising instructions. The observation suggests that the length of viewpoints might serve as a clue because such a viewpoint is explanatory and tends to be long. Introducing such a factor to further improve the proposed method is an interesting future work.

5. CLASSIFICATION EXPERIMENTS

As the number of viewpoints increase, the burden of requesters to analyze the data increase proportionally. Therefore, filtering viewpoints with a classifier would be essential for requesters.

To predict the type of viewpoints, we utilize a simple linear classifier. We use liblinear package [5] specifying L2-regularized logistic regression with hyper-parameters unchanged from the default values. Similar to the document classification problem, an input vector $x_i \in R^m$ represents a viewpoint i , and the output label y_i will be the type of the viewpoint i . We construct input vectors in the following four steps. First, we identify the target object in the original question. For example, if the question is “Is it an instant food?” then the target object is “instant food.” Then in the reason part of a viewpoint, we replace the mentions to the target object with a special symbol T. Second, we identify nouns in the reason part of a viewpoint and replace them as another special symbol O. Third, we concatenate all the resultant strings, after adding the special end of string character to each string. We then extract all maximal substrings from the whole concatenated string [10, 6]. Maximal substrings are essentially every substring found in a document that matters for learning weights of a classifier, and they allow us to obtain the state of art classifier performance [10]. We let these maximal substrings represent the dimensions of the input vectors, so that the j -th dimension of an input vector x_i has a value c if the corresponding maximal substrings occurs c times in the viewpoint i . Finally, we add five more dimensions having binary values, each dimension representing whether or not the answer selected is YES, NO or UNSURE, as well as two indicators that show which of the two text fields are used, representing whether or not the reason part of a viewpoint is input for YES or NO. This setting allows us to detect WRONG label as well.

For the training data, we chose the five drink-related questions about *coffee*, *carbonated drink*, *Japanese tea*, *tea* and *green tea*. For the validation set, the last question about *green tea* is used. The test data consist of the remaining two questions about *seasoning* and *instant food*. As noted in the previous section, we consider instances labeled with PP, NN and PNN as positive instances, and all other instances as negative instances.

The evaluation metrics we use are accuracy (fraction of correctly labeled instances), precision P (fraction of retrieved instances that are relevant), recall R (relevant instances that are retrieved), and their harmonic mean, the F_1 score ($2PR/(P+R)$). As a result of testing, we obtain the classification accuracy of 85.3%. The best F_1 measure we obtain is 74.3% with the precision of 85.8% and the recall of 65.5%. These results tell us that the task is quite difficult to automate given the current size of our corpus.

6. CONCLUSION

In this paper we proposed to collect viewpoints of workers on microtask instructions to help microtask designers reduce ambiguous instructions. In particular, we addressed the problem of choosing useful viewpoints for revising task instructions. Our experiments showed that the proposed logic-based method is comparable to an entropy-based one in the quality of the chosen viewpoints, without requiring each task to be performed by many workers.

7. ACKNOWLEDGMENTS

We are grateful to the project members of Yahoo! Crowdsourcing, including Masashi Nakagawa and Manabu Yamamoto. This work was supported by JST CREST and JSPS KAKENHI Grant Number 25240012 in part.

8. REFERENCES

- [1] S. J. and Lakshminarayanan Subramanian and Ashwin Venkataraman. Reputation-based worker filtering in crowdsourcing. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2492–2500, 2014.
- [2] A. F. and Pavel Kucherbaev and Stefano Tranquillini and Gregorio Convertino. Keep it simple: reward and task design in crowdsourcing. In *Biannual conference of the Italian chapter of SIGCHI, CHIItaly '13, Trento, Italy - September 16 - 20, 2013*, pages 14:1–14:4, 2013.
- [3] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, dec 2008.
- [4] D. E. Difallah, G. Demartini, and P. Cudré-Mauroux. Pick-a-crowd: tell me what you like, and i'll tell you what to do. In *22nd International World Wide Web Conference, WWW'13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 367–374, 2013.
- [5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, jun 2008.
- [6] M. Gallé. The bag-of-repeats representation of documents. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13*, pages 1053–1056, New York, NY, USA, 2013. ACM.
- [7] N. Q. V. Hung, N. T. Tam, N. T. Lam, and K. Aberer. An evaluation of aggregation techniques in crowdsourcing. In *Web Information Systems Engineering - WISE 2013 - 14th International Conference, Nanjing, China, October 13-15, 2013, Proceedings, Part II*, pages 1–15, 2013.
- [8] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.
- [9] M. Melenhorst, M. Menéndez Blanco, and M. Larson. A crowdsourcing procedure for the discovery of non-obvious attributes of social images. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia, CrowdMM '14*, pages 45–48, New York, NY, USA, 2014. ACM.
- [10] D. Okanohara and J. Tsujii. Text categorization with all substring features. In *SIAM International Conference on Data Mining (SDM)*, pages 838–846, 2009.
- [11] G. Tremper. Weakly supervised learning of presupposition relations between verbs. In *Proceedings of the ACL 2010 Student Research Workshop, ACLstudent '10*, pages 97–102, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [12] H. Weisman, J. Berant, I. Szpektor, and I. Dagan. Learning verb inference rules from linguistically-motivated evidence. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12*, pages 194–204, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [13] K. Wiemer-Hastings and P. Wiemer-Hastings. Dp: A detector for presuppositions in survey questions. In *Proceedings of the Sixth Conference on Applied Natural Language Processing, ANLC '00*, pages 90–96, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.