*Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 229–235
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

# Quality Estimation for Partially Subjective Classification Tasks via Crowdsourcing

**Yoshinao Sato, Kouki Miyazawa**
Fairy Devices Inc.
Tokyo, JAPAN
{sato, miyazawa}@fairydevices.jp

**Abstract**

The quality estimation of artifacts generated by creators via crowdsourcing has great significance for the construction of a large-scale data resource. A common approach to this problem is to ask multiple reviewers to evaluate the same artifacts. However, the commonly used majority voting method to aggregate reviewers' evaluations does not work effectively for partially subjective or purely subjective tasks because reviewers' sensitivity and bias of evaluation tend to have a wide variety. To overcome this difficulty, we propose a probabilistic model for subjective classification tasks that incorporates the qualities of artifacts as well as the abilities and biases of creators and reviewers as latent variables to be jointly inferred. We applied this method to the partially subjective task of speech classification into the following four attitudes: agreement, disagreement, stalling, and question. The result shows that the proposed method estimates the quality of speech more effectively than a vote aggregation, measured by correlation with a fine-grained classification by experts.

**Keywords:** crowdsourcing, quality estimation, latent variable model

## 1. Introduction

Crowdsourcing plays an increasingly important role in collecting large-scale data resources in many research fields, including natural language, speech, and image processing. One of the major challenges of crowdsourcing is how to estimate the quality of data. In particular, it is difficult to control the quality of the artifacts generated by anonymous creators via crowdsourcing. An efficient quality estimation method enables us to build a large-scale reliable corpus making the most of available resources. A common approach is to ask multiple reviewers to judge the quality of the artifacts. For quality estimation, we need to consider how to aggregate the answers from reviewers. The conventional method of majority voting, however, does not necessarily work effectively, particularly in the case of a wide range of skills and reliability of the evaluation. While many methods other than simple vote counting have been proposed (Dawid and Skene, 1979; Whitehill et al., 2009; Welinder et al., 2010; Venanzi et al., 2014), most prior research focuses on purely objective tasks, in which we can define a single correct answer. At the opposite extreme, we find diverse responses and little agreement for purely subjective tasks, such as the appraisal of an artwork. Many challenging research topics fall in between these two extremes. In other words, it is significant to study quality estimation for partially subjective tasks. For instance, the classifications of emotion, personality, communication style, and social role can be considered partially subjective. If a task is subjective, the answers from the reviewers tend to disagree because of a wide range of abilities and criteria (Tian and Zhu, 2012). Although several methods have been proposed for the quality estimation for subjective tasks (Baba and Kashima, 2013; Nguyen et al., 2016), it remains to be explored further.

In this paper, we consider a partially subjective classification task. In such a task, a single artifact can definitely be of a particular class or else ambiguous between several classes, but cannot definitely be of several different classes at the same time. In other words, we need to evaluate not only the clarity of being a single class but also the ambiguity between multiple classes. Furthermore, we suppose a two-stage workflow of crowdsourcing that consists of a creation stage and a review stage, as illustrated in Figure 1. In the creation stage, anonymous creators generate artifacts, each of which belongs to one of the given classes. Quality control during this stage is difficult because the creators' abilities and biases have a wide variety. In the review stage, anonymous reviewers evaluate the artifacts to classify each of them as one of the given classes. We need to expect that the evaluations will be divided because the reviewers' abilities and biases are widely varied. Even if we increase the number of reviewers, the significant variance will remain. Therefore, the commonly used majority voting method does not work effectively.

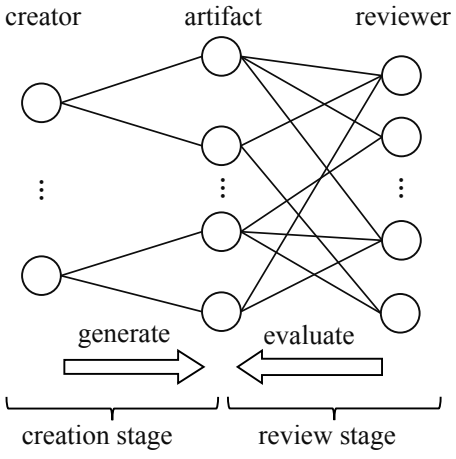


Figure 1: Two-stage workflow of crowdsourcing

To overcome the difficulty of partially subjective classification tasks, we propose a probabilistic model for quality estimation. Our model takes into account the qualities of artifacts, the abilities of creators, and the abilities of re-

viewers as latent variables. The quality estimation is performed by the joint inference of these latent variables. This means our method estimates the "true" quality of artifacts by excluding the subjectivity of crowd workers. We can apply numerical optimization methods, such as the Newton-Raphson method, to the inference. The quality estimation method proposed in (Baba and Kashima, 2013) infers the qualities of artifacts as well as the creators' and reviewers' parameters in a similar way to our method. However, it targets mostly subjective grade rating but not partially subjective classification, which we consider in this paper.

In order to demonstrate the superiority of the proposed method, we applied our method to the classification of speech into four attitudes: agreement, disagreement, stalling, and question. Firstly, we collected a large number of utterances of speech by crowdsourced speakers. Then, the failed recordings were eliminated via another crowdsourced task. After that, each utterance was classified into four attitudes by two or three different reviewers. Finally, we estimated the qualities of the utterances using the proposed method. The effectiveness of the evaluation method was measured by correlation of the estimated quality with a fine-grained evaluation carried out by experts, which is considered to be less subjective.

The result of our experiments shows that the proposed method estimates the qualities of the artifacts more effectively compared to the commonly used vote aggregation method. The proposed method is applicable to collect a large-scale data resource for mutually exclusive classification via crowdsourcing with the two-stage workflow.

# 2. Method

In this paper, we address a data collection procedure via crowdsourcing that consists of two stages: a creation stage and a review stage. Firstly, in the creation state, creators generate artifacts. Then in the reviewer stage, reviewers evaluate the artifacts. Finally, we estimate the quality of the artifacts on the basis of the results of the crowdsourcing. Regarding the task, we consider classifying an artifact $u \in U$ as one of the given classes $i \in I$. Since the classes are mutually exclusive, we suppose the quality of an artifact can be represented by a point on the standard simplex:

$$\left\{ q_{iu} \,\middle|\, i \in I, \sum_{i \in I} q_{iu} = 1, q_{iu} \geq 0 \right\}.$$

This means an artifact can definitely be of a particular class, or else ambiguous between several classes, but it cannot definitely be of several different classes at the same time. For example, $q = (1, 0, \ldots, 0)$ means the artifact definitely belongs to a specific class. On the other hand, $q = (1/|I|, \ldots, 1/|I|)$ indicates the artifact is ambiguous between all classes.

In what follows, we consider two methods for quality estimation: the voting method as a baseline, and the proposed method based on a probabilistic model.

## 2.1. Voting method

As a baseline, we consider the voting method for quality estimation. We count the number of votes in the following way. For each artifact, the class given to the creator who generated it counts as one vote. Besides this, each of the answers given by the reviewers who evaluated it counts as one vote. We get an estimation of the artifact's quality on the standard simplex by dividing the total number of votes.

## 2.2. Probabilistic Model

### 2.2.1. Creation Stage

In the creation stage, each artifact $u$ is generated by a creator $s(u) \in S$ given a class $j(u) \in I$. In other words, each creator $s$ generates a set of utterances $U_s \subseteq U$. We assume that each creator has an ability

$$\{\alpha_{ijs} > 0 \,|\, i \in I, j \in I\}.$$

An artifact with a quality $\{q_{ui}\}_{i \in I}$ is assumed to be generated by a creator $s$ following a Dirichlet distribution with a concentration parameter $\{\alpha_{ij(u)s(u)}\}_{i \in I}$. That is to say, the probability of artifact generation can be written as

$$\begin{aligned} &p\left(\{q_{iu}\}_i \,\middle|\, \{\alpha_{ij(u)s(u)}\}_i\right) \\ &\quad = \mathrm{Dirichlet}\left(\{q_{iu}\}_i; \{\alpha_{ij(u)s(u)}\}_i\right) \\ &\quad = \frac{1}{B(\{\alpha_{ij(u)s(u)}\}_i)} \prod_{i \in I} q_{iu}^{\alpha_{ij(u)s(u)}-1} \quad {}^\forall u \in U. \end{aligned}$$

Here $B$ is the Beta function, which is defined in terms of the gamma function $\Gamma$ by

$$B(\{\alpha_i\}_i) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma\left(\sum_i \alpha_i\right)}.$$

We can say that $a_{js} = \sum_{i \in I} \alpha_{ijs}$ and $m_{ijs} = \alpha_{ijs}/a_{js}$ represent the repeatability and the mean quality of artifacts generated by a creator $s$ given a class $j$.

### 2.2.2. Review Stage

In the review stage, each artifact $u \in U$ is judged by a reviewer $r(u) \in R$ to belong to a class $t_{ur} \in I$. In other words, each reviewer $r$ evaluates a set of artifacts $U_r \subseteq U$ to classify each of them as $\{y_{iur} \,|\, i \in I, u \in U_r\}$, where

$$y_{iur} = \begin{cases} 1 & i = t_{ur} \\ 0 & i \neq t_{ur} \end{cases}.$$

We assume each reviewer has an ability

$$\{\beta_{ir} > 0 \,|\, i \in I\}.$$

Answers given by reviewers $\{y_{iur}\}_{i \in I}$ are assumed to follow a multinomial distribution with a probability parameter

$$\mathrm{softmax}\left(\{\beta_{ir}\eta_{iu}\}_i\right),$$

where $\eta_{iu}$ is defined by $\eta_{iu} = \mathrm{arctanh}\left(q_{iu}\right)$. That is to say, the probability of artifact evaluation is given by

$$\begin{aligned} &p\left(\{y_{iur}\}_i \,|\, \{q_{iu}\}_i, \{\beta_{ir}\}_i\right) \\ &\quad = \mathrm{Mulitonimial}\left(\{y_{iur}\}_i; \mathrm{softmax}\left(\{\beta_{ir}\eta_{iu}\}_i\right)\right) \\ &\quad = \prod_{i \in I} \left[ \frac{\exp\left(\beta_{ir}\eta_{iu}\right)}{\sum_{i' \in I} \exp\left(\beta_{i'r}\eta_{i'u}\right)} \right]^{y_{iur}} \quad {}^\forall u \in U_r, {}^\forall r \in R. \end{aligned}$$

The scale and the relative differences of $\{\beta_{ir}\}_{i \in I}$ represent the sensitivity and the biases of a reviewer $r$ in their evaluation of artifacts, respectively.

#### 2.2.3. Prior Distribution

We introduce prior distributions of the model parameters $\alpha_{ijs}$ and $\beta_{ir}$. We assume a half-Gaussian prior distribution for creators' abilities:

$$p\left(\{\alpha_{ijs}\}_i|\sigma_\alpha\right) \propto \exp\left(-\frac{\left(\sum_i \alpha_{ijs}\right)^2}{2\sigma_\alpha^2}\right) \quad {}^\forall j \in I, {}^\forall s \in S.$$

In a similar way, we assume a half-Gaussian prior distribution for reviewers' abilities:

$$p\left(\{\beta_{ir}\}_i|\sigma_\beta\right) \propto \exp\left(-\frac{\sum_i {\beta_{ir}}^2}{2\sigma_\alpha^2}\right) \quad {}^\forall i \in I, {}^\forall r \in R.$$

In our experiments we set $\sigma_\alpha = \sigma_\beta = \frac{1}{\sqrt{2}} \times 10^2$.

#### 2.2.4. Whole Model

The whole model can be summarized by the graphical model shown in Figure 2.

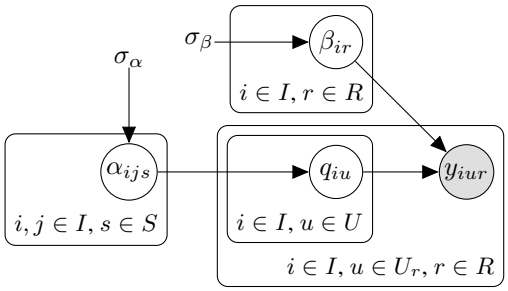


Figure 2: Graphical model of the proposed method

The total log probability is given by

$$\begin{aligned}
&\log p\left(\{y_{iur}\}_{i,u,r}, \{q_{iu}\}_{i,u}, \{\alpha_{ijs}\}_{i,j,s}, \{\beta_{ir}\}_{i,r} \middle| \sigma_\alpha, \sigma_\beta\right) \\
&= \sum_{u \in U} \log p\left(\{q_{iu}\}_i \,\middle|\, \{\alpha_{ij(u)s(u)}\}_i\right) \\
&\quad + \sum_{u \in U_r, r \in R} \log p\left(\{y_{iur}\}_i \,|\, \{q_{iu}\}_i, \{\beta_{ir}\}_i\right) \\
&\quad + \sum_{s \in S, j \in I} \log p\left(\{\alpha_{ijs}\}_i|\sigma_\alpha\right) + \sum_{r \in R} \log p\left(\{\beta_{ir}\}_i|\sigma_\beta\right).
\end{aligned}$$

This can be rewritten as

$$\begin{aligned}
&\sum_{u \in U} \Bigg( \log p\left(\{q_{iu}\}_i \,\middle|\, \{\alpha_{ij(u)s(u)}\}_i\right) \\
&\quad + \sum_{r \in R_u} \log p\left(\{y_{iur}\}_i \,|\, \{q_{iu}\}_i, \{\beta_{ir}\}_i\right) \Bigg) + C_{\alpha,\beta}.
\end{aligned} \tag{1}$$

$C_{\alpha,\beta}$ represents the terms that do not depend on $\{q_{iu}\}_i$, and $R_u = \{r \in R \,|\, u \in U_r\}$ denotes the set of reviewers who evaluate an artifact $u$. We can rewrite the log probability also as

$$\log P_{\mathrm{C}} + \lambda_{\mathrm{R}} \log P_{\mathrm{R}}, \tag{2}$$

where $\log P_C$ and $\log P_R$ are given by

$$\begin{aligned}
\log P_C = \sum_{s \in S, j \in I} \Bigg( &\sum_{u \in U_{js}} \log p\left(\{q_{iu}\}_i \,|\, \{\alpha_{ijs}\}_i\right) \\
&+ \log p\left(\{\alpha_{ijs}\}_i|\sigma_\alpha\right) \Bigg)
\end{aligned} \tag{3}$$

$$\begin{aligned}
\log P_R = \sum_{r \in R} \Bigg( &\sum_{u \in U_r} \log p\left(\{y_{iur}\}_i \,|\, \{q_{iu}\}_i, \{\beta_{ir}\}_i\right) \\
&+ \log p\left(\{\beta_{ir}\}_i|\sigma_\beta\right) \Bigg).
\end{aligned} \tag{4}$$

$U_{js} = \{u \in U \,|\, j = j(u), s = s(u)\}$ denotes the set of artifacts generated by a creator $s$ given a class $j$. In (2), we introduce a parameter $\lambda_{\mathrm{R}}$ that gives weight to the reliability of reviewers than to that of creators. This corresponds to duplicating the observed answers by $\lambda_{\mathrm{R}}$ times.

### 2.3. Inference

We apply a numerical optimization method to infer the latent variables in this study. In order to infer the artifact's qualities $\boldsymbol{q} = \{q_{iu}\}_{i,u}$, the creator's abilities $\boldsymbol{\alpha} = \{\alpha_{ijs}\}_{i,j,s}$, and the reviewer's abilities $\boldsymbol{\beta} = \{\beta_{ir}\}_{i,r}$, we maximize the log likelihood numerically using the following procedure:

1. Initialize $\boldsymbol{q}$.
2. Optimize $\boldsymbol{\alpha}$ with the fixed values of $\boldsymbol{q}$.
3. Optimize $\boldsymbol{\beta}$ with the fixed values of $\boldsymbol{q}$.
4. Optimize $\boldsymbol{q}$ with the fixed values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.
5. Repeat Step 2, 3 and 4 iteratively until the convergence.

We initialize $\boldsymbol{q}$ to the same value as that of the voting method. (3) shows Step 2 can be performed for each creator independently. The optimization of each $\{\alpha_{ijs}\}_i$ is nothing but an estimation of the Dirichlet distribution parameters. In this paper, we apply the optimization method proposed in (Minka, 2000). Similarly, Step 3 can be performed for each reviewer independently, as shown by (4). The optimization of each $\{\beta_{ir}\}_i$ can be formulated as a multinomial logistic regression. We optimize the reviewer's abilities using the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm. Furthermore, (1) indicates that we can decompose Step 4 into independent tasks for each utterance. Since we need to infer an artifact quality on the standard simplex, we apply the trust-region constrained method.

## 3. Dataset

In order to evaluate our method, we built a Japanese speech corpus using crowdsourcing with a two-stage workflow. The task we addressed was a classification of speech into

| attitude | definition |
|---|---|
| Agreement | utterance when the speaker is in favor, is backchanneling, or gives an opinion, often with a falling tone |
| Disagreement | utterance when the speaker is against or dissatisfied, often with a rising tone |
| Stalling | utterance when the speaker is thinking or worried, often with a stretched and leveled tone |
| Question | utterance when the speaker wants to ask, is listening back, or confirms facts, often with a rising tone |

Table 1: Definition of speech attitudes

four attitudes: agreement, disagreement, stalling, and question. Table 1 shows our definition of the attitudes, which were shown to the crowd workers with example audio clips during the data collection.

The speech attitude classification is one of the typical partially subjective tasks. It is known that the intonation curve at the end of accentual or intonational phrases conveys the attitude of the speaker in Japanese (Igarashi et al., 2013). Specific shapes of intonation, including rising, falling, and leveling, are well-associated with attitudes (Venditti, 2005). Nevertheless, there are individual differences in the perceptual boundaries (Kibe et al., 2018). While one reviewer consistently judges artifacts with similar rising intonation curves as members of the disagreement class, another reviewer may classify those same artifacts as members of the question class. In what follows, we describe the details of our corpus.

### 3.1. Recording

Before the main recording, we prepared a qualification round to screen out workers who make low-quality recordings. As a result, 138 speakers went to the main recording step. We asked each of the qualified speakers to read aloud a set of 63 sentences. The sentences assigned to a speaker were randomly chosen from five sets. More specifically, the speakers were asked to read out a given sentence expressing each of the four attitudes. This means that one speaker read aloud 252 utterances. After eliminating invalid recordings, those which contained background noise or speech errors, via another crowdsourcing task, 32,148 utterances remained.

### 3.2. Review

Each of the 20 reviewers was asked to evaluate a set of utterances and to classify them into one of four attitudes. Specifically, we asked the reviewers to gauge the attitudes of the utterances on the basis of the prosodic information but not the linguistic information. The reviewers were not notified of the speaker's attitude at the recording time. We hypothesized that respondents' discrimination between disagreement and question is more subjective than their discrimination between agreement and stalling. Therefore, we assigned three reviewers to utterances in which the speaker had expressed disagreement or a question, while two reviewers were assigned to utterances in which the speaker had expressed agreement or stalling. In total, we collected 80,472 reviews. As for the inter-rater reliability, Krippendorff's $\alpha$ was calculated to be 0.65.

## 4. Experiments

We evaluated the quality of utterances in our corpus by the voting method and the proposed method to investigate their effectiveness. The effectiveness of estimation was measured by correlations with the evaluation by experts on a test set. The result indicates that our method outperforms the commonly used voting method.

### 4.1. Results of Quality Estimation

The distribution of the quality estimated by the voting method and the proposed method with $\lambda_{\mathrm{R}} = 1$, 5, and 10 are shown in Figure 3, 4, 5, and 6, respectively. In addition, Figure 7 illustrates the estimated quality in detail specifically for $\lambda_{\mathrm{B}} = 5$. The diagonal part is the histogram, which is equivalent to that in Figure 5. The upper and lower triangular parts are the scatter and density plots, respectively. As shown in these figures, the voting method yields a sparse discrete distribution. In contrast, the proposed method can estimate the continuous distribution of quality.

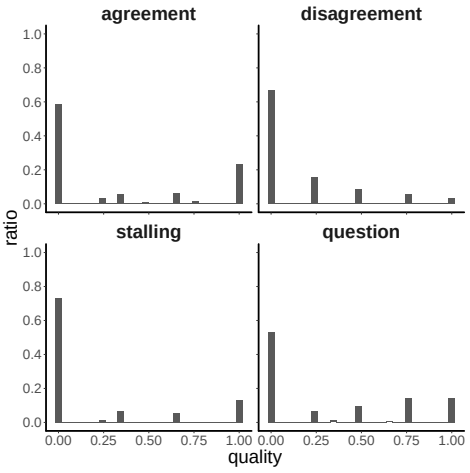


Figure 3: Distribution of the quality estimated by the voting method

### 4.2. Test Set

To analyze the effectiveness of quality estimation quantitatively, we prepared a "less subjective" test set as follows. Firstly, we chose 342 utterances using the random sampling method. In order to balance the quality in the test set, we added weights on the utterances during the sampling. The weight added on each utterance was the inverse of the number of the utterances that had the same quality, eliminating those less than one percent. Since the true quality is unknown, we calculated the weights from the quality estimated by the voting method. Then, we evaluated the quality of the utterance. More specifically, one of the authors classified the test set into the categories shown in Table 2.

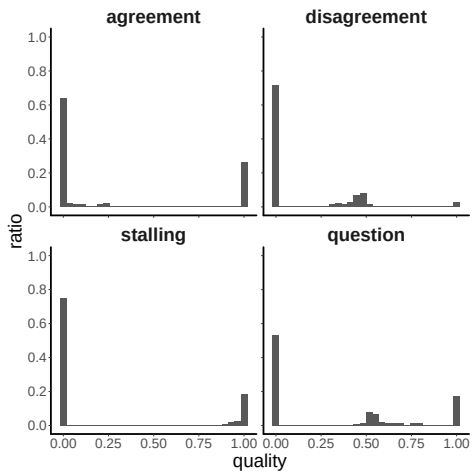


Figure 4: Distribution of the quality estimated by the proposed method with $\lambda_{\mathrm{R}} = 1$

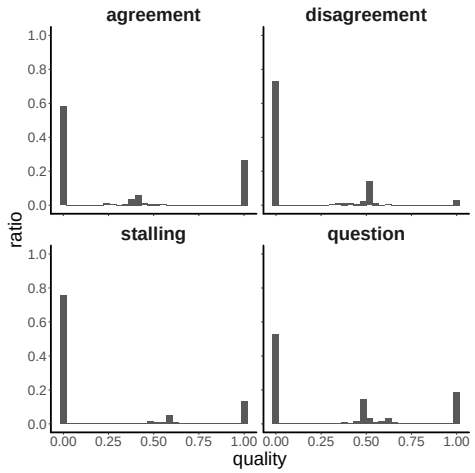


Figure 5: Distribution of the quality estimated by the proposed method with $\lambda_{\mathrm{R}} = 5$

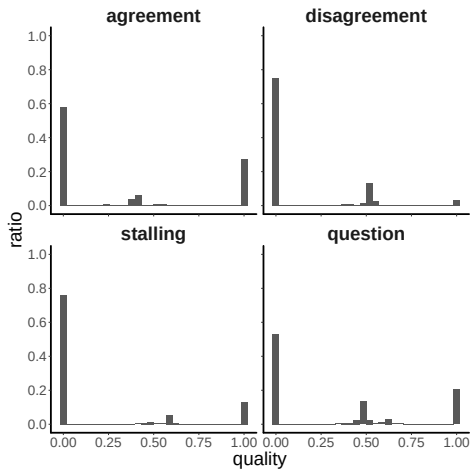


Figure 6: Distribution of the quality estimated by the proposed method with $\lambda_{\mathrm{R}} = 10$

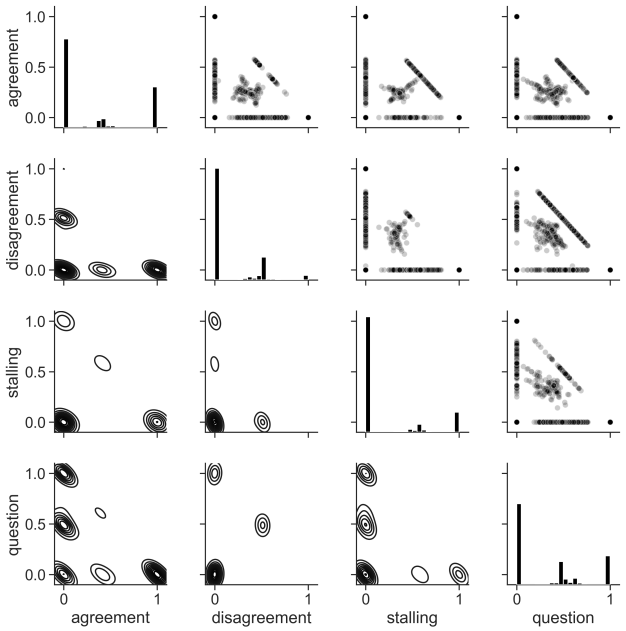


Figure 7: Details of the quality estimated by the proposed method with $\lambda_{\mathrm{R}} = 5$

| grade | definition |
|---|---|
| (5,1,1,1) | definitely A |
| (1,5,1,1) | definitely D |
| (1,1,5,1) | definitely S |
| (1,1,1,5) | definitely Q |
| (4,2,1,1) | probably A, but possibly D |
| ⋮ | ⋮ |
| (1,1,2,4) | probably Q, but possibly S |
| (3,3,1,1) | ambiguous and difficult to distinguish between A and D |
| ⋮ | ⋮ |
| (1,1,3,3) | ambiguous and difficult to distinguish between S and Q |
| (2,2,2,2) | ambiguous between all attitudes |

Table 2: Definition of 23 grades. A, D, S, and Q are the abbreviations of agreement, disagreement, stalling, and question, respectively. The grade is represented as a four-tuple of which elements correspond to the grade of agreement, disagreement, stalling, and question.

As a result, the composition of the test set resulted in what is shown in Figure 8. On the basis that the agreement rate between the authors was higher than 0.85 in a preliminary experiment on a small subset of the corpus, we suppose this evaluation is less subjective than that performed by crowd workers. Nevertheless, it is not feasible to ask crowd workers to perform this task because it demands a high cognitive load as well as expertise.

If we consider grades four and two to be the same as grade three, then the classification is reduced to 11 categories. In addition, the classification can be decomposed into five-grade evaluations of each attitude. In what follows, we call

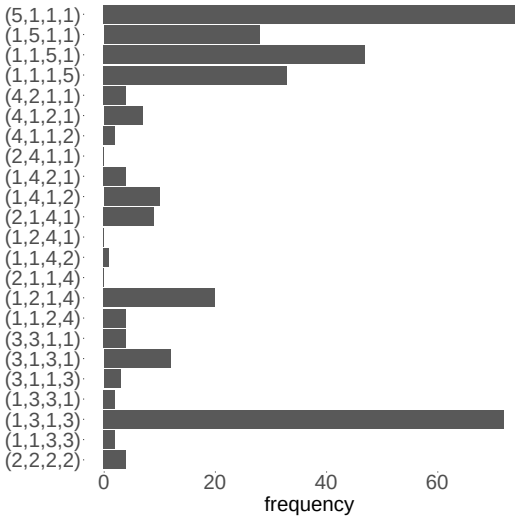


Figure 8: Composition of the test set

the former as a "three-simplex 11-grade" classification and the latter as a "one-dimensional five-grade" estimation.

### 4.3. Evaluation of Quality Estimation

We evaluated the baseline method and the proposed method on the test set. The effectiveness was assessed by comparing the estimated quality to that evaluated by the experts. Specifically, we used the following measures:

- accuracy $p$, Cohen's $\kappa$ (Cohen, 1960) and Aickin's $\alpha$ (Aickin, 1990) in the "three-simplex 11-class" formulation
- Goodman and Kruskal's $\gamma$ (Goodman and Kruskal, 1954) in the "one-dimensional five-grade" formulation

| grade | point on the simplex |
|---|---|
| (5,1,1,1) | $^t$(1 0 0 0) |
| (1,5,1,1) | $^t$(0 1 0 0) |
| (1,1,5,1) | $^t$(0 0 1 0) |
| (1,1,1,5) | $^t$(0 0 0 1) |
| (3,3,1,1) | $^t$(1/2 1/2 0 0) |
| ⋮ | ⋮ |
| (1,1,3,3) | $^t$(0 0 1/2 1/2) |
| (3,3,3,3) | $^t$(1/4 1/4 1/4 1/4) |

Table 3: Correspondence between 11 grades and Voronoi cells

To calculate $p$, $\kappa$, and $\alpha$, we classified the test set into 11 classes on the basis of the estimated quality as follows. Firstly, we mapped each of the 11 grades on to a point on the three-simplex, as shown in Table 3. By using these 11 points as seeds for Voronoi cells, the simplex was partitioned by the Voronoi tessellation. Then, a given utterance with a quality value on the simplex was classified as one of 11 classes with random tie-breaking. In other words, we assigned an utterance to the nearest class on the simplex. The accuracy $p$ is given by the ratio of test data classified by the Voronoi tessellation to the same class as the one by the experts. In the calculation of $\gamma$, we counted a pair that was not tied on the five-grade variable evaluated by experts, but instead on the estimated variable as a discordant one. Subsequently, we got four gamma values with respect to four attitudes and their average. Following the above procedure, we compared the quality estimation methods using the voting method and the proposed method with $\lambda_{\mathrm{R}} = 1$, 5, and 10 to get the results shown in Table 4 and 5. In conclusion, the proposed method outperformed the voting method under all conditions we investigated.

| method | $\lambda_{\mathrm{R}}$ | $p$ | $\kappa$ | $\alpha$ |
|---|---|---|---|---|
| voting | - | 0.48 | 0.40 | 0.43 |
| proposed | 1 | **0.62** | **0.53** | **0.58** |
| proposed | 5 | 0.57 | 0.49 | 0.52 |
| proposed | 10 | 0.55 | 0.47 | 0.50 |

Table 4: Results of the effectiveness evaluation for the "three-simplex 11-class" classification

| method | $\lambda_{\mathrm{R}}$ | $\gamma_{\mathrm{A}}$ | $\gamma_{\mathrm{D}}$ | $\gamma_{\mathrm{S}}$ | $\gamma_{\mathrm{Q}}$ | $\gamma$ |
|---|---|---|---|---|---|---|
| voting | - | 0.82 | 0.45 | 0.56 | 0.64 | 0.62 |
| proposed | 1 | 0.75 | **0.64** | **0.74** | 0.77 | 0.73 |
| proposed | 5 | **0.87** | 0.61 | 0.72 | **0.78** | **0.74** |
| proposed | 10 | **0.87** | 0.60 | 0.72 | 0.74 | 0.73 |

Table 5: Results of the effectiveness evaluation for the "one-dimensional five-grade" evaluation. $\gamma_{\mathrm{A}}$, $\gamma_{\mathrm{D}}$, $\gamma_{\mathrm{S}}$ and $\gamma_{\mathrm{Q}}$ denote the gamma values of agreement, disagreement, stalling, and question, respectively. $\gamma$ denotes their average.

## 5. Conclusion

In this paper, we have addressed quality estimation for partially subjective classification tasks via two-stage crowdsourcing. For this purpose, we have proposed a probabilistic model that enables us to estimate the "true" quality of the artifacts, considering the ability of the creators and the reviewers. Among numerous other partially subjective tasks, we have targeted that of speech attitude classification. We have collected a speech corpus for attitude classification via crowdsourcing with the two-stage workflow. We have estimated the quality of the utterances in our corpus using the commonly used voting method and our method to investigate the effectiveness of both methods, measured by their correlation with a fine-grained quality estimation by experts. The results indicate that the proposed method estimates the quality of artifacts more effectively than does the conventional voting method.

By using an efficient method of quality estimation, we can build a large-scale reliable corpus making full use of finite resources. One way to use such a method is to apply it to an existing corpus with evaluation by reviewers. We can extract a subset of high-quality data from the whole corpus. In

particular, the proposed method is applicable to any corpus with evaluation by reviewers whose task is classification. A generator is not limited to a crowd worker. Different data resources can be regarded as generators. In case no information on the data resources is available, we may suppose all data were created by a single generator. Another way to use a quality estimation method is for building a new corpus. The results of a small-scale preliminary experiment, analyzed by an effective quality estimation method, help us to decide the corpus design. For example, we can estimate an appropriate number of reviewers per artifact. Using our method, we can infer the distribution of quality and the ones of reliability for generators and reviewers. We can design the data collection process based on this information. In these ways, the proposed method of quality estimation enables us to build a large-scale reliable corpus.

Finally, we will discuss some of the future directions of research. The effectiveness of our method with an increased number of reviewers evaluating the same artifact should be investigated. Apart from that, introducing supervision over quality estimation is a possible extension of our method. For example, we can force a part of the artifacts to have given qualities by experts during the inference. This extension may resolve a possible discrepancy between the perceptual scale of the quality and the scale of the estimated values. Moreover, many challenging research topics lie in between purely objective and purely subjective, which we have targeted in this paper, to name a few, the classification of emotion, personality, communication style, and social role.

## 6. Bibliographical References


Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to cohen's kappa. *Biometrics*, pages 293–302.

Baba, Y. and Kashima, H. (2013). Statistical quality estimation for general crowdsourcing tasks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 554–562. ACM.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764.

Igarashi, Y., Nishikawa, K., Tanaka, K., and Mazuka, R. (2013). Phonological theory informs the analysis of intonational exaggeration in japanese infant-directed speech. *The Journal of the Acoustical Society of America*, 134:1283–94.

Kibe, N., Otsuki, T., and Sato, K. (2018). Intonational variations at the end of interrogative sentences in japanese dialects : From the "corpus of japanese dialects". In *Proceedings of the LREC 2018 Special Speech Sessions*.

Minka, T. (2000). *Estimating a Dirichlet Distribution*. Technical report, MIT.

Nguyen, A. T., Halpern, M., Wallace, B. C., and Lease, M. (2016). Probabilistic modeling for crowdsourcing partially-subjective ratings. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

Tian, Y. and Zhu, J. (2012). Learning from crowds in the presence of schools of thought. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 226–234. ACM.

Venanzi, M., Guiver, J., Kazai, G., Kohli, P., and Shokouhi, M. (2014). Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*, pages 155–164. ACM.

Venditti, J. (2005). The j-tobi model of japanese intonation. In Jun Sun-Ah, editor, *Prosodic Typology: The Phonology of Intonation and Phrasing*, vol. 7. Oxford University Press, London.

Welinder, P., Branson, S., Perona, P., and Belongie, S. J. (2010). The multidimensional wisdom of crowds. In *Advances in neural information processing systems*, pages 2424–2432.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R., and Ruvolo, P. L. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pages 2035–2043.