Statistical Quality Estimation for General Crowdsourcing Tasks

Yukino Baba
The University of Tokyo
yukino_baba@mist.i.u-tokyo.ac.jp

Hisashi Kashima
The University of Tokyo
kashima@mist.i.u-tokyo.ac.jp

ABSTRACT

One of the biggest challenges for requesters and platform providers of crowdsourcing is quality control, which is to expect high-quality results from crowd workers who are neither necessarily very capable nor motivated. A common approach to tackle this problem is to introduce redundancy, that is, to request multiple workers to work on the same tasks. For simple multiple-choice tasks, several statistical methods to aggregate the multiple answers have been proposed. However, these methods cannot always be applied to more general tasks with unstructured response formats such as article writing, program coding, and logo designing, which occupy the majority on most crowdsourcing marketplaces. In this paper, we propose an unsupervised statistical quality estimation method for such general crowdsourcing tasks. Our method is based on the two-stage procedure; multiple workers are first requested to work on the same tasks in the creation stage, and then another set of workers review and grade each artifact in the review stage. We model the ability of each author and the bias of each reviewer, and propose a two-stage probabilistic generative model using the graded response model in the item response theory. Experiments using several general crowdsourcing tasks show that our method outperforms popular vote aggregation methods, which implies that our method can deliver high quality results with lower costs.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications—*Data mining*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*

Keywords

crowdsourcing; quality control; human computation

1. INTRODUCTION

Crowdsourcing is a type of online activity of outsourcing specific tasks to a large group of people. With the recent expansion of crowdsourcing platforms such as Amazon Mechanical Turk, one can easily outsource various complex tasks including audio transcription, article writing, language translation, program coding, and graphic designing, as well as simple tasks such as image tagging and Web content categorization. The popularity of crowdsourcing is increasing exponentially in computer science as well, and it has been successfully applied to fields such as natural language processing, computer vision, and human computer interaction (e.g., [1, 2, 18, 19]).

One of the most challenging issues in crowdsourcing research is *quality control* to ensure the quality of crowdsourcing results, because there is no guarantee that all workers have sufficient abilities needed to complete the offered tasks at a satisfactory level of quality. Moreover, it is known that some faithless workers try to get paid as easily as possible, which results in worthless responses. Most crowdsourcing platforms allow requesters to check submitted results, and to reject low-quality results; however, it is not realistic to check all of them manually if their volume is large.

Several approaches to efficient quality control have been proposed. They are roughly categorized into *supervised* and *unsupervised* approaches. Supervised approaches use tasks with known correct answers called *gold standard datasets* to estimate the ability of each worker. For example, each worker is required to qualify before they start the job by passing several tasks selected from the gold standard dataset. They can also be randomly injected into actual tasks in such a way that workers do not recognize them, which allow for ability evaluation of crowd workers to exclude low-quality workers. However, the use of such supervised approaches is limited because of the high cost of preparing the gold standard datasets, or difficulties in determining one unique answer.

Unsupervised quality control methods use *redundancy* instead of the gold standard datasets to ensure work quality; they assign a single task to multiple crowd workers, and aggregate their responses by applying majority voting or more sophisticated statistical aggregation techniques [17]. The statistical quality control methods consider the characteristics of each worker or task, such as the ability of each worker and the difficulty of each task [5, 22, 21]. However, one serious disadvantage of these methods is that most of the existing approaches assume that the response spaces are *structured*. Binary questions (e.g., yes-or-no questions) and

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

ACM 978-1-4503-2174-7/13/08 ...$15.00.

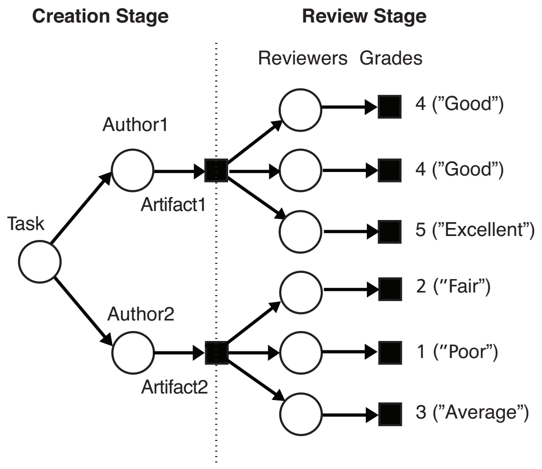


Figure 1: Example of a two-stage workflow comprising a creation stage and a review stage.

multiple-choice questions (e.g., five-point ratings) are typical examples where voting-like strategies work, or we can apply (weighted) averaging to real-valued questions. Unfortunately, these approaches are not applicable for tasks with *unstructured response formats*[1], such as article writing and logo design tasks, where we cannot expect an agreement of two outputs. Most of the crowdsourcing tasks fall into this category. Ipeirotis [8] reported that five of the top twelve Mechanical Turk requesters (based on the total rewards during Jan. 2009–Apr. 2010) posted unstructured response tasks such as content generation, content rewriting, and website feedback. Further, graphic design tasks such as logo design and business card design are quite popular on some specialized crowdsourcing marketplaces such as 99designs and DesignCrowds.

One natural approach to quality estimation of artifacts for general unstructured response tasks is to employ a *two-stage workflow* as shown in Figure 1, consisting of a *creation stage* followed by a *review stage*[2]. In the creation stage, several crowd workers (which we call *authors*) are assigned to several unstructured response tasks. Then, their artifacts proceed to the review stage, where each of them is reviewed by multiple crowd workers (called *reviewers*). The review tasks are usually casted as multiple-choice questions (such as 'Excellent,' 'Good,' 'Average,' 'Fair,' and 'Poor'). Although it is quite difficult to estimate the quality of the artifacts directly from themselves, introducing the review stage enables us to indirectly estimate the quality from the review scores, and to distinguish high-quality results from the others. For example, Zaidan and Callison-Burch [24] applied the two-stage workflow; however, their approaches are supervised so that they require extensive domain knowledge including feature representation of artifacts and gold standard scores.

In this paper, we propose an unsupervised statistical method to estimate the quality of artifacts of general unstructured response tasks using the framework of the two-stage workflow. We introduce a two-stage generative model (Figure 2). The creation stage models a generative process of the true artifact quality, where both the ability and the task-dependent performance of an author affect the quality of an artifact. The review stage models the generative process of the grade labels given by reviewers, where each reviewer first determines a latent quality score for a given artifact based on their bias and contextual preference, and then the observed grade label is generated through the graded response model [16] used in the item response theory [20]. The true artifact quality and the model parameters are estimated using the maximum a posteriori (MAP) inference. The proposed algorithm consists of simple iterations of a closed-form update and a convex optimization.

We conduct experiments using logo designing tasks, image description tasks, and language translation tasks on a commercial crowdsourcing platform. Our method outperforms the other methods, including the majority voting and an ordinal label aggregation method [15] (which is an extension of the well-known method proposed by Dawid and Skene [5]) with a small number of reviewers and a moderate number of authors. The result implies that our method can deliver high-quality results with lower costs, because the number of involved workers directly affects the total monetary and time costs.

In summary, this paper makes three main contributions:

1. We address an *unsupervised* statistical quality estimation problem for *general crowdsourcing tasks with unstructured response formats* such as article writing, program coding, and logo designing (Section 2).

2. We introduce a *two-stage generative model* for the general crowdsourcing processes consisting of the creation stage and the review stage (Section 3). In both stages, the ability or bias of workers are incorporated in the model, and the true quality and the review scores of artifacts are affected by them.

3. We devise an efficient iterative algorithm which performs a MAP inference of the true artifact quality as well as the other parameters (Section 4).

2. PROBLEM SETTING

We start with a formulation of the workflow of general crowdsourcing tasks with unstructured response formats consisting of two-stages: the creation stage and the review stage (Figure 1). Further, we describe the quality estimation problem of general crowdsourcing tasks.

Let us assume that there is a set of general crowdsourcing tasks $\mathcal{T}$ (such as logo designing and content generation), and let $\mathcal{A}_t$ denote a set of crowd authors assigned to a task $t \in \mathcal{T}$. In the creation stage, each author $a \in \mathcal{A}_t$ creates an artifact for a task t. We denote the (unknown) quality of the artifact by $q_{t,a} \in \mathbb{R}$. In the review stage, a set of crowd reviewers $R_{t,a}$ is assigned to evaluate the quality of the artifact created by author a for task t. The evaluation by a reviewer $r \in R_{t,a}$ is given as a grade label $g_{t,a}^{(r)}$ from a set of grade labels $\mathcal{D} = \{1, 2, \cdots, n\}$.

Our goal is to estimate the set of the true qualities of the artifacts $\{q_{t,a}\}_{t\in\mathcal{T},a\in\mathcal{A}_t}$, given the set of the observed grade labels $\{g_{t,a}^{(r)}\}_{t\in\mathcal{T},a\in\mathcal{A}_t,r\in\mathcal{R}_{t,a}}$.

In practice, the set of authors and reviewers may overlap, and some reviewers possibly give good grades to their own

[1] Lin et al. [11] considered tasks with somewhat unstructured formats; however, they still assume that two output instances agree.

[2] They are called by different names in literature (e.g., [12]).

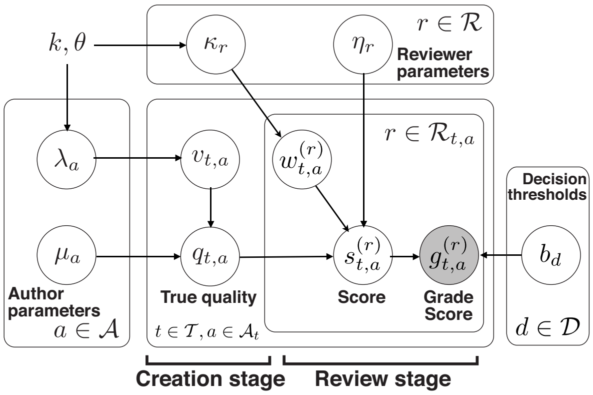


Figure 2: Graphical model of our proposed two-stage model. $\mu_a \in \mathbb{R}$ denotes the ability of the author $a \in \mathcal{A}$, and $1/\lambda_a \in \mathbb{R}^+$ denotes the variance of the artifact-specific noise $v_{t,a} \in \mathbb{R}$ for the pair of the task $t \in \mathcal{T}$, and the author a. The true quality $q_{t,a}$ of the output is given as the sum of μ_a and $v_{t,a}$. $\eta_r \in \mathbb{R}$ denotes the evaluation bias of the reviewer $r \in \mathcal{R}$, and $1/\kappa_r \in \mathbb{R}^+$ denotes a variance of the contextual preference $w^{(r)}_{t,a} \in \mathbb{R}$ for the artifact created by the author a for the task t. The quality score $s^{(r)}_{t,a}$ is the sum of η_r, $w^{(r)}_{t,a}$, and the true quality $q_{t,a}$, which results in the observed grade $g^{(r)}_{t,a} \in \{1, 2, \ldots, n\}$ through the graded response model with threshold parameters $\{b_d\}_d$. k and θ are hyper-parameters.

output. However, we assume that we can exclude such workers with some identifiers; in other words, the sets of authors and reviewers are distinct.

3. TWO-STAGE MODELING OF GENERAL CROWDSOURCING TASKS

To estimate the true quality $q_{t,a}$ of the artifact created by author a for task t, we introduce a two-stage generative model, where the first stage models the generation of the artifact of quality $q_{t,a}$, and the second stage models the generation of the grade label $g^{(r)}_{t,a}$ given by reviewer r to the artifact. Figure 2 shows the graphical model of our grade label generation process.

3.1 Creation Stage

We assume that an author with a higher ability creates higher-quality artifacts on average; hence, each author $a \in$ has ability $\mu_a \in \mathbb{R}$. We also assume that the performance of an author on each task varies according to the type and instance of the task. Considering language translation tasks as an example, even an author with a low general translation skill might sometimes produce high quality translations for sentences related to information technologies, if he is knowledgeable about information technologies. We model such variety depending on the combination of task t and author a as the noise $v_{t,a} \in \mathbb{R}$. We assume that the noise $v_{t,a}$ follows a Gaussian distribution with zero mean and a variance of $1/\lambda_a$ (i.e., a precision of λ_a); that is,

$$v_{t,a} \sim \mathcal{N}(v_{t,a} \mid 0, 1/\lambda_a) = \sqrt{\frac{\lambda_a}{2\pi}} \exp\left(-\frac{\lambda_a v_{t,a}^2}{2}\right). \tag{1}$$

Note that each author a has their own λ_a.

At the end of the creation stage, the quality of the artifact $q_{t,a} \in \mathbb{R}$ is given as the sum of the general ability and the artifact-specific variation, namely,

$$q_{t,a} = \mu_a + v_{t,a}.$$

3.2 Review Stage

In the review stage, we assume that each reviewer r has a base bias $\eta_r \in \mathbb{R}$, assuming that a reviewer with a lower bias tends to give lower grades to the given artifacts, and one with a higher bias gives higher grades. We also incorporate the contextual preferences of reviewers, for example, some reviewers might prefer short sentences to long sentences. We model such preferences as the noise depending on a pair of output and a reviewer denoted by $w^{(r)}_{t,a} \in \mathbb{R}$. We assume that $w^{(r)}_{t,a}$ follows a Gaussian distribution with zero mean and a variance of $1/\kappa_r$ (i.e., a precision of κ_r); that is,

$$w^{(r)}_{t,a} \sim \mathcal{N}\left(w^{(r)}_{t,a} \mid 0, 1/\kappa_r\right). \tag{2}$$

Note that each reviewer r has their own κ_r. When reviewer $r \in \mathcal{R}_{t,a}$ evaluates the output of author a for task t, the reviewer first estimates the (latent) quality score $s^{(r)}_{t,a} \in \mathbb{R}$ of the output, which is given as the sum of the true quality of an artifact, $q_{t,a}$, the reviewer's bias η_r, and contextual preference $w^{(r)}_{t,a}$, namely,

$$s^{(r)}_{t,a} = q_{t,q} + \eta_r + w^{(r)}_{t,a}. \tag{3}$$

Finally, since the final grade label $g^{(r)}_{t,a}$ is a discrete value depending on the quality score, we apply $\Pr[g^{(r)}_{t,a} = d \mid s^{(r)}_{t,a}]$, which is the conditional probability of selecting $d \in \mathcal{D}$ given the quality score $s^{(r)}_{t,a}$. For modeling $\Pr[g^{(r)}_{t,a} = d \mid s^{(r)}_{t,a}]$, we adopt the graded response model (GRM) [16] (Figure 3), which is a standard model of the graded responses of subjects in the item response theory (IRT) [20]. In the GRM, the conditional probability of a graded response is decomposed by using $n-1$ binary response models, namely,

$$\begin{aligned}\mathrm{GRM}\left(g^{(r)}_{t,a} = d \mid s^{(r)}_{t,a}\right) &= \Pr[g^{(r)}_{t,a} = d \mid s^{(r)}_{t,a}] \\ &= \Pr[g^{(r)}_{t,a} > d-1 \mid s^{(r)}_{t,a}] - \Pr[g^{(r)}_{t,a} > d \mid s^{(r)}_{t,a}],\end{aligned}$$

where $\Pr[g^{(r)}_{t,a} > 0 \mid s^{(r)}_{t,a}] = 1$ and $\Pr[g^{(r)}_{t,a} > n \mid s^{(r)}_{t,a}] = 0$. There are several possible choices for the binary response models, and we adopt the Rasch model [14], which is one of the simplest models, given as

$$\Pr[g^{(r)}_{t,a} > d \mid s^{(r)}_{t,a}] = \sigma\left(s^{(r)}_{t,a} - b_d\right) = \frac{1}{1 + \exp\left(-(s^{(r)}_{t,a} - b_d)\right)},$$

where σ is the sigmoid function, and $\{b_d\}_d$ are threshold parameters. Finally, our grade label generation model is

$$\mathrm{GRM}\left(g^{(r)}_{t,a} = d \mid s^{(r)}_{t,a}\right) = \sigma(s^{(r)}_{t,a} - b_{d-1}) - \sigma(s^{(r)}_{t,a} - b_d).$$

For simplicity, we set the thresholds $(b_1, b_2, \cdots, b_{n-1}) = (1, 2, \cdots, n-1)$ in our implementation, because it had no significant effect on the performance.

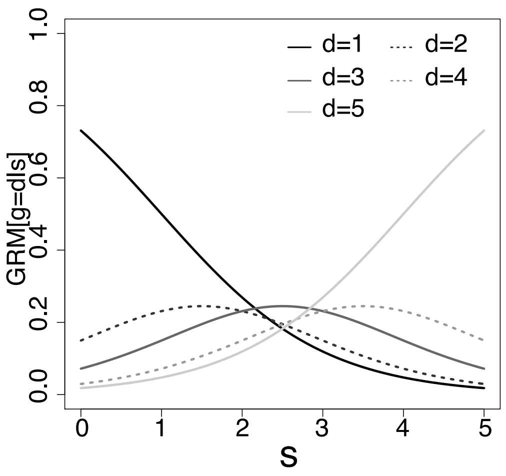


Figure 3: Example of a probability density function of the graded response model (GRM), which models the probability of a graded response g given a latent score s, where a set of grade labels $\mathcal{D} = \{1, 2, 3, 4, 5\}$ and the threshold parameters $(b_1, b_2, b_3, b_4) = (1, 2, 3, 4)$

4. QUALITY ESTIMATION

Based on the two-stage crowdsourcing model introduced in the previous section, we introduce our approach that uses the maximum a posteriori (MAP) inference to estimate the artifact quality as well as the other parameters such as the abilities of workers. Our algorithm consists of simple iterations of two optimization steps: one is a convex optimization problem, and the solution of the other step is given in closed forms.

4.1 Priors

We introduce prior distributions on the model parameters to apply the MAP inference. In the creation stage, we assume that the prior for worker ability is given as a Gaussian distribution with zero mean and a variance of 1; that is,

$$\mu_a \sim \mathcal{N}(\mu_a \mid 0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\mu_a^2}{2}\right).$$

Since the precision parameter λ_a in the artifact-specific noise (1) must be positive, we assume a Gamma prior,

$$\lambda_a \sim \text{Gamma}(\lambda_a \mid k, \theta) = \frac{1}{\theta^k} \frac{1}{\Gamma(k)} \lambda_a^{k-1} \exp\left(-\frac{\lambda_a}{\theta}\right),$$

where k and θ are hyperparameters.

Similarly, in the review stage, we give the prior for the bias of each worker as a Gaussian distribution,

$$\eta_r \sim \mathcal{N}(\eta_r \mid 0, 1),$$

and assume a Gamma prior on the precision parameter κ_r for the contextual preference (2); that is,

$$\kappa_r \sim \text{Gamma}(\kappa_r \mid k, \theta),$$

where k and θ are hyperparameters.

4.2 Objective Function

The total likelihood L is a function of $\{\mu_a\}_a$, $\{\lambda_a\}_a$, $\{\eta_r\}_r$, $\{\kappa_r\}_r$, $\{v_{t,a}\}_{t,a}$, and $\{w_{t,a}^{(r)}\}_{t,a,r}$ given as

$$\begin{aligned} L = &\prod_a \mathcal{N}(\mu_a \mid 0, 1)\, \text{Gamma}(\lambda_a \mid k, \theta) \\ &\times \prod_r \mathcal{N}(\eta_r \mid 0, 1)\, \text{Gamma}(\kappa_r \mid k, \theta) \\ &\times \prod_{t \in \mathcal{T}} \prod_{a \in \mathcal{A}_t} \mathcal{N}(v_{t,a} \mid 0, 1/\lambda_a) \\ &\times \prod_{t \in \mathcal{T}} \prod_{a \in \mathcal{A}_t} \prod_{r \in \mathcal{R}_{t,a}} \mathcal{N}\left(w_{t,a}^{(r)} \mid 0, 1/\kappa_r\right) \text{GRM}\left(g_{t,a}^{(r)} \mid s_{t,a}^{(r)}\right) \end{aligned}$$

where $s_{t,a}^{(r)}$ is defined in Eq. (3). Its logarithm is given as

$$\begin{aligned} \log L = &\sum_a \left(-\frac{\mu_a^2}{2} + (k-1)\log \lambda_a - \frac{\lambda_a}{\theta}\right) \\ &+ \sum_r \left(-\frac{\eta_r^2}{2} + (k-1)\log \kappa_r - \frac{\kappa_r}{\theta}\right) \\ &- \sum_{t \in \mathcal{T}} \sum_{a \in \mathcal{A}_t} \frac{\lambda_a}{2} v_{t,a}^2 - \sum_{t \in \mathcal{T}} \sum_{a \in \mathcal{A}_t} \sum_{r \in \mathcal{R}_{t,a}} \Bigg(\frac{\kappa_r}{2} w_{t,a}^{(r)\,2} \\ &\qquad + \log\left(\sigma(s_{t,a}^{(r)} - b_{g_{t,a}^{(r)}-1}) - \sigma(s_{t,a}^{(r)} - b_{g_{t,a}^{(r)}})\right)\Bigg) \\ &+ \sum_{a \in \mathcal{A}} \frac{|\mathcal{T}_a|}{2} \log \lambda_a + \sum_{r \in \mathcal{R}} \frac{|\mathcal{U}_r|}{2} \log \kappa_r, \end{aligned} \tag{4}$$

where $\mathcal{T}_a$ denotes the set of tasks done by author a, and $\mathcal{U}_r$ is the set of task-author pairs evaluated by reviewer r.

4.3 Optimization

Our strategy to optimize the objective function is to split the parameters into the set of $\{\lambda_a\}_a$ and $\{\kappa_r\}_r$ and the set of the other parameters, $\{\mu_a\}_a$, $\{\eta_r\}_r$, $\{v_{t,a}\}_{t,a}$, and $\{w_{t,a}^{(r)}\}_{t,a,r}$, because the optimal solutions with respect to $\{\lambda_a\}_a$ and $\{\kappa_r\}_r$ are given in closed forms. The optimization problem with respect to the other parameters is a convex programming problem; hence, we can apply standard nonlinear optimization methods such as the gradient ascent and the Newton-Rhapson method. These facts naturally give the following iterative optimization procedure.

1. Set initial parameters (to zeros)
2. Maximize the objective function (4) w.r.t. $\{\lambda_a\}_a$ and $\{\kappa_r\}_r$ (using Eqs. (5))
3. Maximize the objective function (4) w.r.t. $\{\mu_a\}_a$, $\{\eta_r\}_r$, $\{v_{t,a}\}_{t,a}$, and $\{w_{t,a}^{(r)}\}_{t,a,r}$ (using a numerical optimization method)
4. If the solution has not yet converged, go to Step 2

The closed form solution of Step 2 is given as

$$\lambda_a = \frac{k - 1 + \frac{|\mathcal{T}_a|}{2}}{\frac{1}{\theta} + \sum_{t \in \mathcal{T}_a} \frac{v_{t,a}}{2}} \quad , \quad \kappa_r = \frac{k - 1 + \frac{|\mathcal{U}_r|}{2}}{\frac{1}{\theta} + \sum_{t,a \in \mathcal{U}_r} \frac{w_{t,a}^{(r)}}{2}}, \tag{5}$$

if $k > 1 - \frac{|\mathcal{T}_a|}{2}$ and $k > 1 - \frac{|\mathcal{U}_r|}{2}$. Note that this condition is always satisfied if $|\mathcal{T}_a| > 1$ and $|\mathcal{U}_r| > 1$, or $k > 1/2$.

In our implementation, we employed a simple gradient ascent to solve the convex optimization problem in Step 3.

Table 1: Statistics about the datasets in the creation stage

	#tasks	#unique authors	Avg. #authors per task	Avg. #tasks per author	Reward for each task	#all created artifacts
Logo Designing	34	47	20.9	15.1	N/A	710
Image Description	20	20	10.0	10.0	$0.04	200
Language Translation	20	17	9.5	11.2	$0.09	190

Table 2: Statistics about the datasets in the review stage

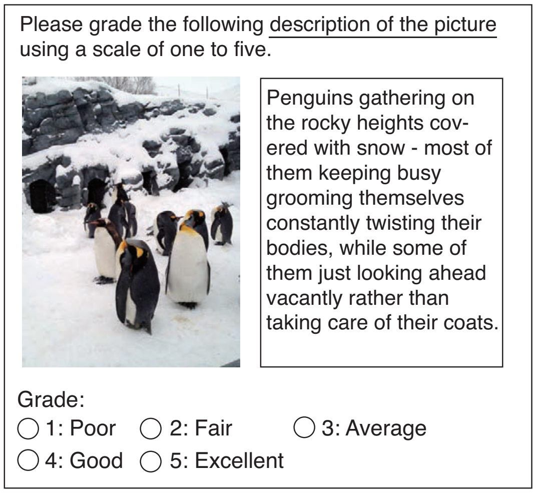
	#reviewed artifacts	#unique reviewers	Avg. #reviewers per artifact	Avg. #artifacts per reviewer	Reward for each artifact	#all obtained grade labels
Logo Designing	710	155	25.0	114.5	$0.01	17750
Image Description	200	87	25.0	52.0	$0.01	5000
Language Translation	190	71	24.4	65.2	$0.01	4630

5. EXPERIMENTS

To evaluate our proposed two-stage model, we prepared three tasks, logo designing (Section 5.1.1), image description (Section 5.1.2), and English-to-Japanese translation (Section 5.1.3), and posted them on a commercial crowdsourcing service. Using the artifacts obtained in the creation stage, we posted review tasks for each artifact. We compared the accuracies of the qualities estimated by our two-stage model with those by three other methods (Section 5.2).

5.1 Datasets

We built our datasetsusing the Lancers crowdsourcing marketplace [3] . Tables 1 and 2 give their general statistics.



Please grade the following description of the picture using a scale of one to five.

Penguins gathering on the rocky heights covered with snow - most of them keeping busy grooming themselves constantly twisting their bodies, while some of them just looking ahead vacantly rather than taking care of their coats.

Grade:
○ 1: Poor ○ 2: Fair ○ 3: Average
○ 4: Good ○ 5: Excellent

Figure 4: Example of a review task for an image description

5.1.1 Logo Designing Task

Graphic design is a typical example of unstructured response format tasks. Design tasks usually take the form of contest. A requester chooses the most preferable design from ones submitted by crowd designers. Only the winner gets the prize, and the others are not paid (or paid only a small amount of money).

We collected the data from 34 (already closed) logo design contests from Lancers, and used the submitted logos as the artifacts in the creation stage. We posted evaluation tasks asking workers to give five-point ratings to the artifacts; each evaluation task includes 10 logo designs.

5.1.2 Image Description Task

Textual descriptions for images are useful resources for enhancing image search accuracy, or to help visually impaired people understand the content of a picture. Generating image description is a typical example of a problem that is relatively easy for a human but very difficult for a computer.

We requested tasks of writing a description of a picture within 140 Japanese characters. We randomly selected 20 pictures from the SBU Captioned Photo Dataset [13]. Each author was asked to complete one or more batch of tasks comprising 10 randomly selected pictures. After the completion of the description task, we posted tasks of reviewing the submitted descriptions. Figure 4 shows an example of the review tasks. Reviewers were instructed to review the description in terms of both adequacy and fluency, and assign a five-point grade to each description.

5.1.3 Language Translation Task

Language translation is one of the most common tasks in crowdsourcing marketplaces, and several research efforts (e.g. [24]) attempt to collect high-quality translations from non-professional translators using crowdsourcing.

We posted sentence translation tasks from English to Japanese, and then posted grading tasks for each submitted translation. We selected 20 English sentences from the Japanese-English Bilingual Corpus of Wikipedia's Kyoto Article[4] . We made batches, each of which consisted of ten randomly selected sentences, and one of these batches was assigned to each author.

[3] http://www.lancers.jp

[4] http://alaginrc.nict.go.jp/WikiCorpus/index_E.html

In the review stage, we requested crowdsourcing workers to review each of the translated Japanese sentences and to assign a five-point grade to it. We asked the reviewers to take into account the fluency and adequacy of each sentence when assigning a grade.

5.2 Comparison Methods

The review stage makes it possible to apply label aggregation methods to quality estimation for unstructured response format tasks. The existing label aggregation methods for multiple-choice questions can be applied to the collected grade labels $\{g_{t,a}^{(r)}\}_{t,a,r}$. We compare our proposed two-stage model with two aggregation methods: majority voting and the modified Dawid-Skene model [15]. Majority voting is a simple method to aggregate the labels; however, it shows a good performance on several crowdsourcing tasks [18, 9]. The label aggregation proposed by Dawid and Skene [5] has been successfully applied to various crowdsourcing applications. They give a generative model of labels created by each worker with their own ability, where the ability of a worker is represented as a confusion matrix which represents the conditional probability of an observed label given a true label. The true labels and the confusion matrices are estimated by using the EM algorithm. Since we focus on ordinal scores in this paper, we use the modified version proposed by Raykar et al. [15] (which we call '*ordinal Dawid-Skene*').

The major differences between our proposed two-stage model and the other two competing methods are summarized as follows. To estimate the quality of the artifact by author a for task t, majority voting only uses $\{g_{t,a}^{(r)}\}_r$, the graded labels limited to the given artifact. On the other hand, the Dawid-Skene model and our two-stage model exploit $\{g_{t,a}^{(r)}\}_{t,a,r}$, *all* the graded labels in the dataset. Furthermore, our two-stage model incorporates both the abilities of the workers and the biases of the reviewers in contrast with the Dawid-Skene model which only considers the abilities of the reviewers.

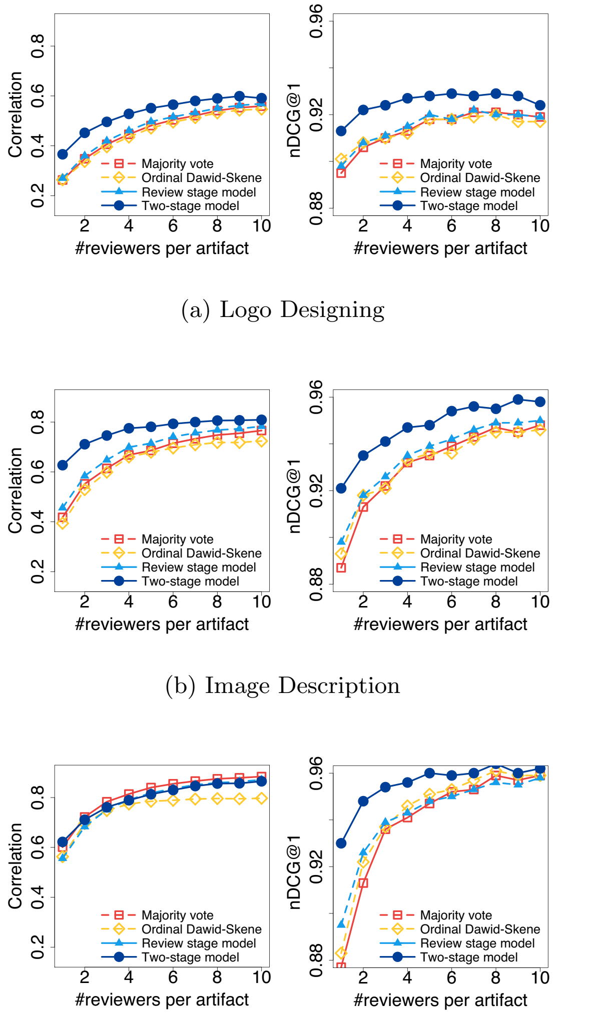
To evaluate the advantage of introducing the creation stage, we also tested our two-stage model without the creation stage (which we call '*review stage model*'). Concretely, we fixed the parameters in the creation stage at the prior means, i.e., $\mu_a = 0$ and $\lambda_a = k\theta$ for each author a.

5.3 Evaluation Methodology

We calculated the correlation coefficients between the estimated artifact quality scores and the ground truth grades. We also evaluated nDCG@1, which is defined as the ratio of the true quality of the estimated best artifact to that of the true best artifact. because we are often interested in finding the best artifact for each task. Since we could not know the "ground truths," we simulated the ground truth scores using majority voting with sufficiently many labels. Concretely, we used a small part of the collected grade labels for estimation, and used the others for simulating the ground truth scores. This is supported by the results of Snow et al. [18], where they suggested majority voting with ten or more nonexpert worker is on par with that with experts for various NLP tasks. For example, we collected 25 grade labels for each artifact in the image description dataset as in Table 2.

We also investigated the impact on the estimation accuracy by the number of authors assigned to each task and the number of reviewers assigned to each artifact. We varied the number of reviewers for each artifact from one to ten. Similarly, we varied the number of authors from one to its maximum value for each task.

For statistical testing, we sampled 100 subsets of the data for each (n, m) pair, where m denotes the number of authors and n denotes the number of reviewers, and performed the Wilcoxon signed rank test. We set $k = 16$ and $\theta = 0.5$ throughout the experiments.



(a) Logo Designing

(b) Image Description

(c) Language Translation

Figure 5: Correlation and nDCG@1 between estimated quality scores and ground truth scores along with the number of reviewers per artifact. In most cases, the proposed two-stage model outperforms the other three baselines especially with small numbers of reviewers.

5.4 Results

Table 3 and Figure 5 show the correlations and nDCG@1 between estimated artifact scores and ground truth scores for each number of reviewers (ranging from one to ten). In most cases, our proposed two-stage model achieved statisti-

Table 3: Averages and standard deviations of correlations and nDCG@1 between estimated quality scores and ground truth scores along with the number of reviewers per artifact. Statistically significant ($p < 0.05$) winners by the Wilcoxon signed rank test are bold-faced. In most of the cases, the proposed two-stage model outperforms the other three baselines in both the quality of overall ranking and finding the best artifact.

	correlation				nDCG@1			
#reviewers per artifact	1	3	5	10	1	3	5	10
Logo Designing								
Majority vote	0.263 (±0.032)	0.406 (±0.027)	0.481 (±0.023)	0.559 (±0.017)	0.895 (±0.012)	0.910 (±0.011)	0.918 (±0.010)	0.919 (±0.011)
Ordinal Dawid-Skene	0.264 (±0.034)	0.396 (±0.028)	0.471 (±0.025)	0.547 (±0.018)	0.901 (±0.010)	0.910 (±0.012)	0.918 (±0.010)	0.917 (±0.010)
Review Stage model	0.270 (±0.033)	0.420 (±0.028)	0.497 (±0.023)	0.569 (±0.017)	0.898 (±0.011)	0.911 (±0.011)	0.920 (±0.010)	0.919 (±0.012)
Two-Stage model	**0.366** (±0.035)	**0.496** (±0.023)	**0.551** (±0.019)	**0.591** (±0.017)	**0.913** (±0.013)	**0.924** (±0.010)	**0.928** (±0.010)	**0.924** (±0.010)
Image Description								
Majority vote	0.418 (±0.063)	0.614 (±0.041)	0.686 (±0.027)	0.766 (±0.020)	0.887 (±0.023)	0.922 (±0.017)	0.935 (±0.017)	0.948 (±0.013)
Ordinal Dawid-Skene	0.394 (±0.058)	0.599 (±0.043)	0.679 (±0.029)	0.724 (±0.023)	0.893 (±0.023)	0.921 (±0.016)	0.936 (±0.015)	0.946 (±0.014)
Review Stage model	0.456 (±0.062)	0.648 (±0.038)	0.715 (±0.027)	0.783 (±0.020)	0.898 (±0.020)	0.926 (±0.016)	0.939 (±0.015)	0.950 (±0.012)
Two-Stage model	**0.627** (±0.054)	**0.746** (±0.025)	**0.781** (±0.021)	**0.809** (±0.015)	**0.921** (±0.019)	**0.941** (±0.014)	**0.948** (±0.015)	**0.958** (±0.010)
Language Translation								
Majority vote	0.601 (±0.041)	**0.783** (±0.025)	**0.840** (±0.016)	**0.884** (±0.011)	0.877 (±0.026)	0.936 (±0.018)	0.947 (±0.015)	0.959 (±0.011)
Ordinal Dawid-Skene	0.563 (±0.043)	0.748 (±0.027)	0.785 (±0.020)	0.797 (±0.015)	0.883 (±0.025)	0.937 (±0.020)	0.951 (±0.016)	0.959 (±0.012)
Review Stage model	0.556 (±0.043)	0.751 (±0.023)	0.818 (±0.017)	0.871 (±0.012)	0.895 (±0.026)	0.939 (±0.017)	0.948 (±0.014)	0.958 (±0.011)
Two-Stage model	**0.622** (±0.032)	0.761 (±0.018)	0.813 (±0.016)	0.865 (±0.012)	**0.930** (±0.023)	**0.954** (±0.013)	**0.960** (±0.010)	0.962 (±0.012)

cally significant higher performance over the other methods. In particular, when the number of reviewers is small, our method showed large improvements. It is notable that our model performed better even in such cases where we had only one reviewer and therefore the voting-like strategies do not work. This is because our model incorporates the creation stage with the ability parameters of authors for making the most of available information. The difference of the performance between the two-stage model and the review stage model (i.e., a two-stage model with fixed author parameters) shows the benefit of modeling the variance of author abilities.

Only in the language translation task, the simple majority voting performed the best in terms of the correlation measure. This is partly explained by Figure 6 showing the distribution of the correlations between the scores given by each reviewer and the ground truths. While the reviewer abilities widely distribute in the design task and the description task, those in the translation task skew to large positive values, which implies the majority of the reviewers are reliable.

The overall improvements in the nDCG@1 measure show the proposed model is good at finding the best artifact, which is a desirable feature in crowdsourcing-contest-style scenarios, where tasks are highly heterogeneous with unstructured response formats, and domain knowledge such as features and ground truths are not available.

Finally, we investigate the impact of changing the number of authors assigned to each task. Figure 7 shows the average correlations and nDCG@1 with varying number of authors for each task (with the number of reviewers fixed at three). Again, in most of the cases, our proposed two-stage model with moderate numbers of authors outperformed the others. Note that nDCG@1 degrades with increased number of authors due to its definition, since it becomes more difficult to find the best one as the number of artifacts increases.

In summary, in terms of both the quality of overall ranking and the accuracy of finding the best artifact, we verified the effectiveness of our two-stage model, especially with a small number of reviewers and a moderate number of authors.

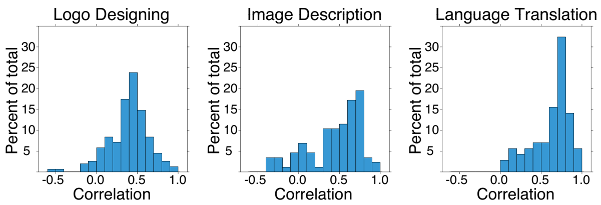
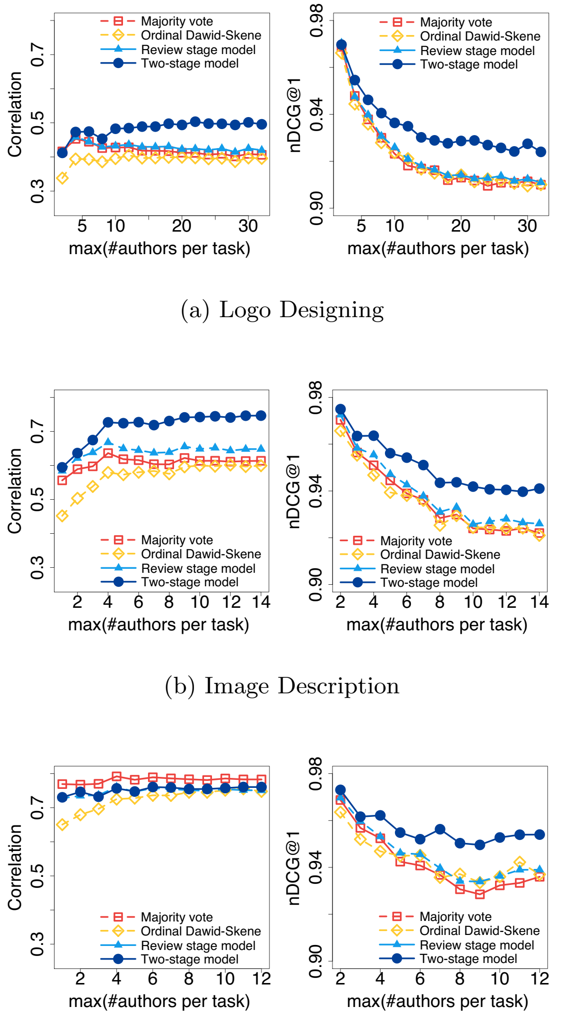


Figure 6: Distributions of the correlations between each reviewer's scores and ground truth scores. The distribution in the language translation task skews to large positive values.



(a) Logo Designing

(b) Image Description

(c) Language Translation

Figure 7: Correlation and nDCG@1 between estimated quality scores and ground truth scores along with the number of authors per task. The number of reviewers per artifact is fixed at three. In most cases, the proposed two-stage model outperforms the other three baselines with moderate numbers of authors per task.

6. RELATED WORK

A number of unsupervised methods were proposed for quality control of structured response format tasks. Most of them guarantee the quality by assigning each task to multiple workers and by aggregating redundant answers. The simplest way to aggregate answers is taking majority votes, and it is used in various NLP tasks [18] and information retrieval tasks [9]. Inspired by the seminal work of Dawid and Skene [5] who modeled the generative process of the answers of workers by introducing their ability parameters, many more sophisticated aggregation methods were proposed. To name a few, Whitehill et al. also included the difficulty of the task in the model [22], and Welinder et al. proposed a model considering workers compatibility for each task [21].

Lin et al. proposed an aggregation method for the tasks that deals with unstructured format responses [11]; however, they targeted only the tasks where each answer possibly agrees to one of the others (e.g., arithmetic problems), and cannot be applied to such tasks we considered in this paper.

Other domain specific quality estimation method for unstructured response format tasks were studied, for example, for language translation tasks [24]. Although they consider the similar problem as ours, their approach is specialized for translation task, and requires extensive domain knowledge including feature representation of artifacts and gold standard scores for employing a supervised learning approach.

While our work addresses the parallel procedure, Dai et al. [4] proposed a quality control method for an iterative improvement procedure [12]. They offered review tasks of comparing artifacts before and after the improvement and proceeded to the next improvement task for a better enhancement.

Reviewing processes of scientific papers and commercial products have similar form to crowdsourcing with unstructured output formats. There are several attempts to obtain appropriate review scores by correcting reviewer-dependent biases [7, 10]; however, the existing models do not include the creation process which we consider in this paper.

7. CONCLUSION

In this paper, we proposed an unsupervised statistical method to estimate the quality of the artifacts for a general crowdsourcing tasks with unstructured response formats. We proposed a two-stage model consisting of the creation stage, where multiple authors create their outputs for same tasks, and the review stage, where another set of workers review and grade the outputs. Our model introduced both the ability of each author and the bias of each reviewer, and modeled the process of grade label selection by reviewers by using the graded response model in the item response theory. We also proposed a simple iterative algorithm for the MAP inference of the true quality and model parameters. Experimental results showed the advantage of our two-stage model compared with some existing label aggregation methods, especially when limited numbers of reviewers and authors are available, which implies that the proposed method can deliver high-quality crowdsourcing results with lower costs.

Finally, we mention some possible future work. We employed the absolute scoring in the review stage, that is, we asked each reviewer to assign grade labels. Instead, we can also use relative scoring such as pairwise ranking [3] by asking which one of two given artifacts is better. Design of the review tasks is also an important open question. Although we requested the reviewers to evaluate randomly chosen artifacts at once, showing artifacts from the same task may be an alternative method. Active selection of tasks and workers is also an important direction to pursue [17, 6, 23].

8. ACKNOWLEDGMENTS

This research was supported by the Funding Program for World-Leading Innovative R&D on Science and Technology (FIRST Program).

9. REFERENCES

[1] M. Bernstein, G. Little, R. Miller, B. Hartmann, M. Ackerman, D. Karger, D. Crowell, and K. Panovich. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM Symposium on User Interface Software and Technology (UIST)*, 2010.

[2] J. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. Miller, R. Miller, A. Tatarowicz, B. White, S. White, et al. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM Symposium on User Interface Software and Technology (UIST)*, 2010.

[3] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM)*, 2013.

[4] P. Dai, Mausam, and D. S. Weld. Artificial intelligence for artificial artificial intelligence. In *Proceedings of the 25th Conference on Artificial Intelligence (AAAI)*, 2011.

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

[6] P. Donmez, J. G. Carbonell, and J. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.

[7] P. A. Flach, S. Spiegler, B. Golénia, S. Price, J. Guiver, R. Herbrich, T. Graepel, and M. J. Zaki. Novel tools to streamline the conference review process: experiences from SIGKDD'09. *SIGKDD Explorations*, 11(2):63–67, 2010.

[8] P. G. Ipeirotis. Analyzing the Amazon Mechanical Turk marketplace. *ACM XRDS*, 17(2), 2010.

[9] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *Proceedings of the 34th Anual ACM SIGIR Conference (SIGIR)*, 2011.

[10] H. W. Lauw, E.-P. Lim, and K. Wang. Summarizing review scores of "unequal" reviewers. In *Proceedings of the 7th SIAM International Conference on Data Mining (SDM)*, 2007.

[11] C. Lin, M. Mausam, and D. Weld. Crowdsourcing control: Moving beyond multiple choice. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

[12] G. Little, L. Chilton, M. Goldman, and R. Miller. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2010.

[13] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24*, 2011.

[14] G. Rasch. *Probabilistic models for some intelligence and attainment tests*. Denmarks Paedagogiske, 1960.

[15] V. C. Raykar and S. Yu. Ranking annotators for crowdsourced labeling tasks. In *Advances in Neural Information Processing 24*, 2011.

[16] F. Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 1969.

[17] V. Sheng, F. Provost, and P. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge discovery and Data Mining (KDD)*, 2008.

[18] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

[19] A. Sorokin and D. Forsyth. Utility data annotation with Amazon Mechanical Turk. In *Proceedings of the 1st IEEE Workshop on Internet Vision*, 2008.

[20] W. van der Linden and R. Hambleton. *Handbook of modern item response theory*. Springer, 1996.

[21] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23*, 2010.

[22] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, 2009.

[23] Y. Yan, R. Rosales, G. Fung, and J. G. Dy. Active learning from crowds. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011.

[24] O. F. Zaidan and C. Callison-Burch. Crowdsourcing translation: professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, 2011.