Crowdordering

Toshiko Matsui[1], Yukino Baba[1], Toshihiro Kamishima[2], and Hisashi Kashima[1,3]

[1] The University of Tokyo
matsui@sr3.t.u-tokyo.ac.jp,
{yukino_baba,kashima}@mist.i.u-tokyo.ac.jp
[2] National Institute of Advanced Industrial Science and Technology (AIST)
mail@kamishima.net
[3] JST PRESTO

Abstract. Crowdsourcing is a promising solution to problems that are difficult for computers, but relatively easy for humans. One of the biggest challenges in crowdsourcing is quality control, since high quality results cannot be expected from crowdworkers who are not necessarily very capable or motivated. Several statistical crowdsourcing quality control methods for binary and multinomial questions have been proposed. In this paper, we consider tasks where crowdworkers are asked to arrange multiple items in the correct order. We propose a probabilistic generative model of crowd answers by extending a distance-based order model to incorporate worker ability, and propose an efficient estimation algorithm. Experiments using real crowdsourced datasets show the advantage of the proposed method over a baseline method.

1 Introduction

Crowdsourcing offers online marketplaces where specific tasks can be outsourced to a large group of people. With the recent expansion of the use of crowdsourcing platforms, such as Amazon Mechanical Turk, various professional and non-professional tasks, including audio transcription, article writing, language translation, program coding, and graphic designing, can now easily be outsourced. The popularity of crowdsourcing is increasing exponentially in computer science as well, and researchers exploit it as an efficient and inexpensive way to process a large number of tasks that humans can perform much more easily than computers, such as image annotation and web content categorization. Crowdsourcing has been successfully applied to such fields as natural language processing, computer vision, and human computer interaction [1–4].

One of the most challenging problems in crowdsourcing research is achieving *quality control* to ensure the quality of crowdsourcing results, because there is no guarantee that the ability of all workers is sufficient to complete the offered tasks at a satisfactory level of quality. Moreover, it is known that some untrustworthy workers try to receive remuneration while expending as little effort as possible, which results in outputs of no value. Most crowdsourcing platforms allow requesters to check the submitted results and to reject low-quality results; however, if their volume is large, realistically, they cannot all be checked manually.

One popular approach to the quality control problem is to use tasks with known correct answers to evaluate the ability of each worker. This approach has been implemented

V.S. Tseng et al. (Eds.): PAKDD 2014, Part II, LNAI 8444, pp. 336–347, 2014.

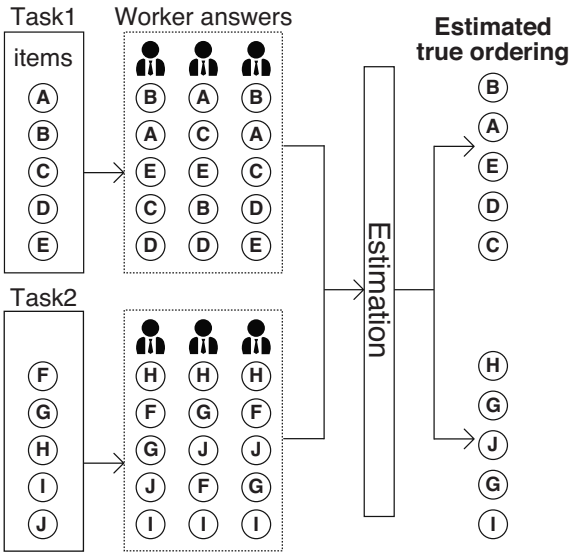


Fig. 1. Overview of quality control problem for item ordering tasks in crowdsourcing. The objective is to estimate true ordering of given items from crowd-generated answers for each item ordering task.

on several commercial crowdsourcing platforms such as CrowdFlower; however, its usage is limited because of the high cost of preparing the correct answers or the difficulty of determining one unique answer. Another promising approach is to introduce *redundancy*. A single task is assigned to multiple workers, and their responses are aggregated by majority voting [5] or more sophisticated statistical aggregation techniques that consider the characteristics of each worker or task, such as the ability of each worker and the difficulty of each task [6–8].

In most existing approaches, it is assumed that the tasks are binary questions to which binary answers (e.g., "yes" or "no") are expected, or multiple-choice questions. Only a few methods have been proposed that extend the applicability of the aggregation-based quality control approach to more general crowdsourcing tasks [9]. Following the the same line, we consider *item ordering tasks*, where workers are asked to arrange multiple items in the correct order. Item ordering tasks, typical examples of which are the ranking of web search results and ordering of items in a to-do list in according to their dependencies [10], are frequently posted on crowdsourcing sites.

In this paper, we propose an aggregation-based statistical quality control method for item ordering tasks. We model the generative process of a worker response (i.e., an ordering of items) using a distance-based probabilistic ordering model [11]. The ability of each worker is naturally incorporated into the concentration parameter of the distance-based model. We also present an effective algorithm for estimating the true ordering, which is particularly efficient because the Spearman distance [12] is employed as the distance measure between two different orderings of items.

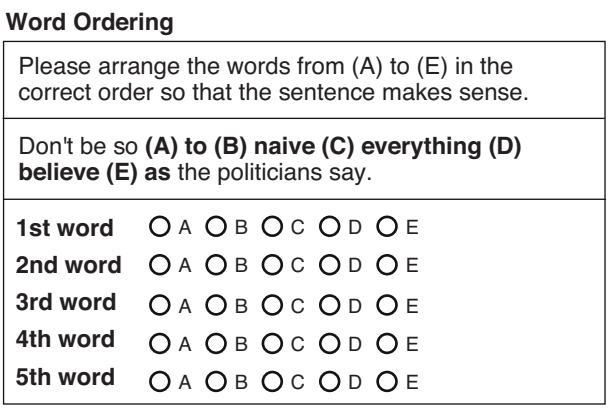


Fig. 2. Example of a word ordering task. The correct answer is (B)(E)(A)(D)(C).

It should be noted that Chen et al. also proposed a quality control method for item ordering tasks [13] based on a pairwise ranking model; however, their method focuses on finding the correct ordering of a single (large) set of items, whereas our method focuses on solving multiple different (relatively small) ordering tasks simultaneously. Additionally, since their method is based on pairwise comparisons, it is not always suitable for tasks where more than two items are needed to determine their correct order. Fig. 2 shows an example of such a task.

We describe our experiments in which word and sentence ordering tasks were posted on a commercial crowdsourcing marketplace. We compare our quality control method to an aggregation method that does not consider the abilities of workers. The experimental results show that our method achieves answers that are more accurate than those of baseline method.

In summary, this paper makes three main contributions:

1. We address the quality control problem for a set of item ordering tasks (Section 2).
2. We propose a generative model of worker responses to item ordering tasks that extend a distance-based probabilistic ordering model to incorporate the ability of each worker (Section 3).
3. We introduce an efficient algorithm to estimate the true ordering from multiple worker responses (Section 4).

2 Crowdsourcing Quality Control for Item Ordering Tasks

We first define the crowdsourcing quality control problem related to item ordering tasks, where each ordering task requires crowdworkers to place given items in the correct order. We then present a model for aggregating the answers collected from multiple workers to obtain answers that are more accurate.

Let us assume I ordering tasks, whose i-th task has M_i items to be ordered. The true order is represented as a *rank vector* $\boldsymbol{\pi}_i = (\pi_{i,1}, \pi_{i,2}, \ldots, \pi_{i,M_i})$, where $\pi_{i,j}$

indicates the position of item j of task i in the true order of the items of M_i [11]. For example, for a task with five items indexed as $1, 2, 3, 4$, and 5, whose true order is given as $(2, 4, 1, 3, 5)$, the true rank vector is $(3, 1, 4, 2, 5)$. Note that $\boldsymbol{\pi}_i$ is a permutation of $(1, 2, \ldots, M_i)$.

We resort to crowdsourcing to obtain estimates for the true rank vectors. It is assumed that a total of K crowdworkers is employed. In the following, $\mathcal{I}^{(k)}$ denotes the indices of tasks on which the k-th worker works, and $\mathcal{K}_i$ denotes the indices of the workers who work on the i-th task. $\boldsymbol{\pi}_i^{(k)} = (\pi_{i,1}^{(k)}, \pi_{i,2}^{(k)}, \ldots, \pi_{i,M_i}^{(k)})$ denotes the rank vector that the k-th worker gives to the i-th item ordering task.

Our goal is to estimate the true rank vectors $\{\boldsymbol{\pi}_i\}_{i \in \{1,2,\ldots,I\}}$ given the (unreliable) rank vectors $\{\boldsymbol{\pi}_i^{(k)}\}_{k \in \{1,2,\ldots,K\}, i \in \mathcal{I}^{(k)}}$ collected using crowdsourcing.

3 Model

To resolve the issue of the aggregation problem of the crowd-generated answers to item ordering tasks, we present a statistical model of the generative process of worker responses, so that we apply statistical inference to estimate the true order from the observed responses.

3.1 Distance-Based Model for Orders

We first review the probabilistic ordering model on which our generative model of crowdworker responses is based. We chose a distance-based model [11] from several variations of the ordering models. A distance-based model gives the probability of a rank vector $\tilde{\boldsymbol{\pi}}$, given a modal order $\boldsymbol{\pi}$ and a concentration parameter λ, namely,

$$\Pr[\tilde{\boldsymbol{\pi}} \mid \boldsymbol{\pi}, \lambda] = \frac{1}{Z(\lambda)} \exp\left(-\lambda d(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi})\right),$$

where $d(\cdot, \cdot)$ denotes a distance between two rank vectors, and $Z(\lambda)$ is a normalizing constant given as

$$Z(\lambda) = \sum_{\tilde{\boldsymbol{\pi}}} \exp\left(-\lambda d(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi})\right).$$

Specifically, we employ the Euclidean distance (also referred to as the *Spearman distance* in the ranking model literature) due to its convenience for deriving an effective parameter estimation method, which will be described later. The distance-based model in which the Spearman distance is applied is called the Mallows θ model [12].

3.2 Extension of the Distance-Based Model for the Crowdsourcing Setting

In crowdsourcing, some workers may have sufficient abilities to provide accurate orders, while some are unskilled and often submit wrong orders. To capture such worker characteristics, we incorporate the worker dependent concentration parameters into the

distance-based ordering model. Namely, it is assumed that the k-th worker has his/her own personal concentration parameter $\lambda^{(k)}$, and the generative model for the worker is then given as

$$\Pr[\tilde{\boldsymbol{\pi}} \mid \boldsymbol{\pi}, \lambda^{(k)}] = \frac{1}{Z(\lambda^{(k)})} \exp\left(-\lambda^{(k)} d(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi})\right).$$

In this model, the answer of a worker who has a high concentration parameter $\lambda^{(k)}$ is likely to be an accurate order whose distance from the true order (i.e., the modal order $\boldsymbol{\pi}$) is small. Therefore, we can interpret the personal concentration parameter $\lambda^{(k)}$ as the ability parameter of the k-th worker.

4 Estimation

Based on the distance-based crowd-ordering model introduced in the previous section, we introduce a maximum likelihood estimation method to obtain estimates for the true rank vectors as well as the worker ability parameters. Our strategy for optimization is to repeat two optimization steps: optimizing the true rank vector and optimizing the worker ability.

4.1 Objective Function

We apply the maximum likelihood estimation to estimate the true rank vector $\{\boldsymbol{\pi}_i\}_i$ and the worker ability parameters $\{\lambda^{(k)}\}_k$, given the crowd-generated rank vectors $\{\boldsymbol{\pi}_i^{(k)}\}_{i,k}$. The objective function for the maximization problem is the log-likelihood function L, given as

$$\begin{aligned} L(\{\lambda^{(k)}\}_k, \{\boldsymbol{\pi}_i\}_i) &= \sum_k \sum_{i \in \mathcal{I}^{(k)}} \log \frac{1}{Z(\lambda^{(k)})} \exp\left(-\lambda^{(k)} d(\boldsymbol{\pi}_i^{(k)}, \boldsymbol{\pi}_i)\right) \\ &= -\sum_k \sum_{i \in \mathcal{I}^{(k)}} \left\{ \lambda^{(k)} d(\boldsymbol{\pi}_i^{(k)}, \boldsymbol{\pi}_i) + \log \sum_{\tilde{\boldsymbol{\pi}}} \exp\left(-\lambda^{(k)} d(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi}_i)\right) \right\}. \quad (1) \end{aligned}$$

4.2 Optimization

Our strategy for optimizing the objective function (1) w.r.t. $\{\lambda^{(k)}\}_k$ and $\{\boldsymbol{\pi}_i\}_i$ is to repeat the two optimization steps, that w.r.t. $\{\lambda^{(k)}\}_k$ and that w.r.t. $\{\boldsymbol{\pi}_i\}_i$. Since L is *not* a convex function, and therefore, its solution depends on the initial parameters, we start with the solution assuming all workers have equal abilities, specifically, $\lambda^{(k)} = \lambda^{(\ell)}$ for an arbitrary pair of k and ℓ.

One major virtue of our model is that the optimization problem is decomposable with respect to each worker and task, that is, each small optimization problem solved at each iteration step depends always on one single variable (a worker ability or a mode order), so that the computational cost linearly is dependent on the numbers of workers and tasks.

Optimization w.r.t. True Rank Vectors. Given that all the worker ability parameters $\{\lambda^{(k)}\}_k$ are fixed, the true rank vectors $\{\boldsymbol{\pi}_i\}_i$ are obtained by maximizing the first term of the objective function (1). Optimization with respect to $\{\boldsymbol{\pi}_i\}_i$ is a combinatorial optimization problem that is often computationally hard to solve; however, we are able to solve it efficiently by employing the Spearman distance as the distance measure $d(\cdot, \cdot)$.

The optimal true rank vector $\boldsymbol{\pi}_i$ for task i is given as follows[1]. First, for each item $m(= 1, \ldots, M_i)$, we calculate a *weighted rank* $w_{i,m}$, which is a weighted mean of the ranks given by workers weighted by the worker abilities,

$$w_{i,m} = \frac{1}{|\mathcal{K}_i|} \sum_{k \in \mathcal{K}_i} \lambda^{(k)} \pi^{(k)}_{i,m}.$$

The maximum likelihood estimator of the true item ordering is given by sorting the items by $w_{i,1}, w_{i,2}, \ldots, w_{i,M_i}$ in ascending order. It should be noted that each $\{\boldsymbol{\pi}_i\}_i$ is obtained independently of the others.

Optimization w.r.t. Worker Ability Parameters. Optimization with respect to the worker ability parameters $\lambda^{(k)}$ with fixed true rank vectors $\{\boldsymbol{\pi}_i\}_i$ is performed by numerical optimization. The objective function (1) is represented as the sum of the different objective functions $\{J^{(k)}\}_k$, where $J^{(k)}$ for each k is defined as

$$J^{(k)}(\lambda^{(k)}) = \lambda^{(k)} d(\boldsymbol{\pi}^{(k)}_i, \boldsymbol{\pi}_i) + \log \sum_{\tilde{\boldsymbol{\pi}}} \exp\left(-\lambda^{(k)} d(\tilde{\boldsymbol{\pi}}, \boldsymbol{\pi}_i)\right).$$

Noting that $J^{(k)}(\lambda^{(k)})$ depends only on $\lambda^{(k)}$, we can consider K independent optimization problem with only one variable.

Since only a single variable function $J^{(k)}(\lambda^{(k)})$ needs to be considered to optimize $\lambda^{(k)}$, the optimization is easily performed by applying a standard optimization method. In the experiments, we employed a simple gradient descent method.

5 Experiments

We collected two crowdsourced datasets, one for word ordering tasks, and the other for sentence ordering tasks. We experimentally evaluated the advantages of our model as compared to a baseline method.

5.1 Datasets

We collected two datasets using Lancers[2], which is a general purpose crowdsourcing marketplace. Table 1 gives the general statistics of the datasets.

Word Ordering. Word ordering is a task whose objective is to order given English words into a grammatically correct sentence. The word ordering problem can be a

[1] Due to the space limitation, we omit the proof of the optimality.

[2] `http://lancers.jp`

Table 1. Statistics about the datasets

	#tasks	#workers	Avg. #items per task	Reward for each task	#all obtained orderings
Word ordering	20	15	5.2	\$0.05	300
Sentence ordering	13	15	5.1	\$0.07	195

subproblem of machine translation between languages with different grammatical word ordering, such as English to Japanese translation. Although several methods have been proposed to solve this ordering problem [14], computer programs still cannot easily perform this task. However, humans, especially the native speakers of the target language, can skillfully perform the word ordering tasks. The workers were given an English sentence with five or six randomly shuffled words, and asked to correct the order of the words. An example of the task is given in Fig. 2. Since we had the correct order of each sentence as the ground truth, we could evaluate the accuracies of our estimation results.

Sentence Ordering. Sentence ordering is a task in which given sentences are ordered such that the aligned texts logically make sense. It emulates several tasks that we presume are posted in crowdsourcing marketplaces, for example, to revise a piece of writing such that its focal point is emphasized more clearly ,or ordering items in a to-do list by their dependencies [10]. In each sentence ordering task, a paragraph consisting of five or six sentences whose order was permuted was presented to the workers, and they were requested to arrange the sentences correctly. Fig. 3 shows an example of the sentence ordering task.

5.2 Results

We applied our method to the two crowd-generated datasets, and calculated the Spearman distance (i.e., the squared error) between each estimated rank vector and the ground truth rank vector. We also tested a baseline method that does not consider the workers' ability. Concretely, we fixed the worker ability parameter $\lambda^{(k)} = 1$ for all workers k, and then optimized the objective function (1) with respect only to $\{\boldsymbol{\pi}_i\}_i$. It should be noted that our proposed method uses the solution of this baseline method as the initial parameters.

The number of workers involved in each task directly affects the monetary cost of posting tasks to an actual crowdsourcing marketplace. In order to investigate the impact on the estimation accuracy engendered by the number of workers assigned to each task, we randomly selected n (ranging from 3 to 15) workers from the all workers for each task, and only used the responses of the selected workers for the estimation. We examined the averaged estimation errors of 50 trials. The results are shown in Fig. 4.

In the word ordering task, our proposed method drastically reduced the estimation error of the baseline method when the number of workers assigned to each task was more than four. It is worth mentioning that the averaged squared error of our method was

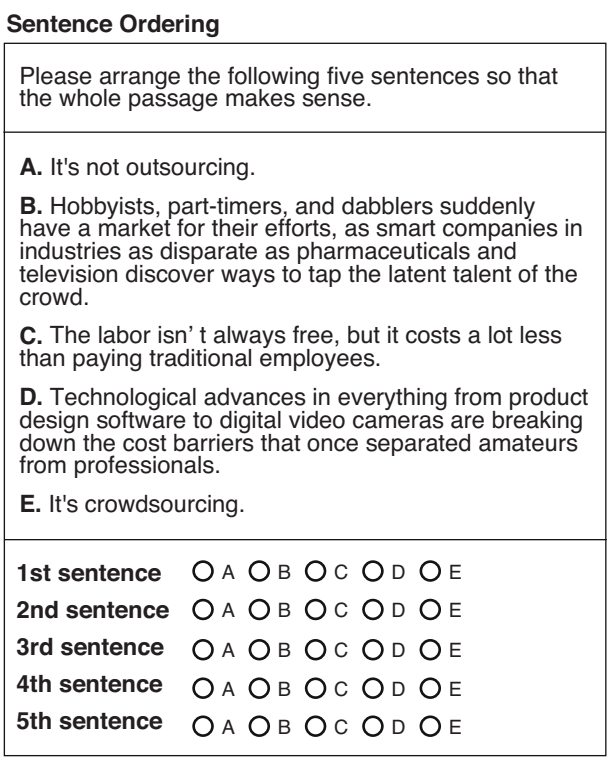
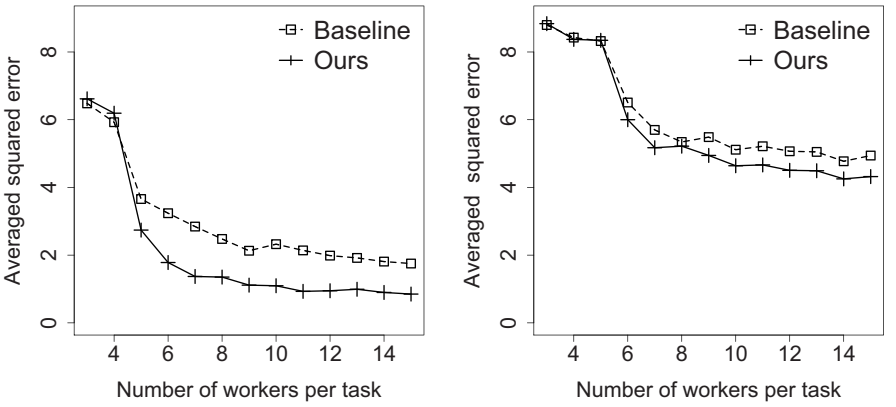


Fig. 3. Example of a sentence ordering task. The passage in this example is from *The Rise of Crowdsourcing* by Jeff Howe. The correct answer is (D)(B)(C)(A)(E).

only 0.902 when all the collected responses were used, while the squared error easily reached 2. When the order of a pair of items that were adjacent in the correct order were incorrectly estimated, the squared error was 2. For example, a rank vector was estimated as $(2, 1, 3, 4, 5)$, when the correct one was $(1, 2, 3, 4, 5)$. This result implies that our method reduces the number of such errors by approximately half.

Our method outperformed the baseline method in the sentence ordering task as well, when the number of workers assigned to each task was more than five. Since the sentence ordering task is generally more difficult than the word ordering task, the averaged estimation errors of both the proposed and baseline methods in the sentence ordering task increased as compared with those in the word ordering task. The best result achieved by our method was a squared error of 4.25, which is relatively large; however, considering the expected squared errors when using random guessing is 21.2, it can be said the result is acceptable. In addition to the Spearman distance, we compared our method and the baseline method in three different measures shown in Table 2. The results in all the measures demonstrated the performance improvement of our method in both the word ordering and sentence ordering tasks.



(a) Word ordering (b) Sentence ordering

Fig. 4. Accuracy evaluation of estimated orders comparing the proposed method and the baseline method. Averaged squared errors between estimated orders and ground truth orders along with the number of workers per task are shown. In both the word ordering and sentence ordering tasks, the proposed method outperforms the baseline method in most cases.

Fig. 5 shows the relations between the estimated worker ability parameters $\{\lambda^{(k)}\}_k$ and the averaged squared errors of each worker (against the ground truths). These results show that the true worker ability (i.e., the worker error versus the ground truths) certainly varies from person to person, and that the proposed method gives higher weights to superior workers, which explains its improved performance. In fact, the estimated worker abilities and the worker errors showed strong negative correlations of -0.853 and -0.695 for the word ordering tasks and the sentence ordering tasks, respectively.

Finally, we mention the scalability of our proposed method. Generally, estimated orders show convergence after five to ten iterations. The ability parameters for good workers require more iterations than those for inferior workers. As discussed before, the complexity of each iteration depends linearly on the numbers of workers and tasks.

In summary, we verified that the proposed method shows clear advantages as compared to the baseline method for estimating the correct orders in both the word ordering and sentence ordering tasks. We also confirmed that the proposed method precisely estimates worker ability.

6 Related Work

One of the fundamental challenges in crowdsourcing is controlling the quality of the obtained data. Crowdworkers are rarely trained and they do not necessarily have adequate ability to complete the tasks [3]. There also exist large differences between the

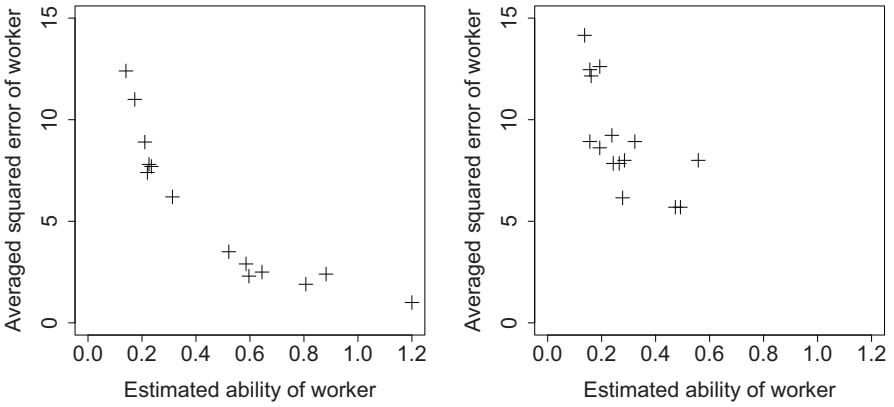
Table 2. Evaluation of estimated orders in several measures. *Error rate* is the fraction of tasks where the estimated order did not exactly match with the ground truth order, *Hamming distance* counts the number of items whose position is different from the ground truth, and *Kendall distance* counts the number of item pairs who are in the opposite order of the ground truth. Average Hamming distance and Kendall distance of task are shown. Number of workers per task is fixed to 15. The results in all of these measures clearly indicate that the proposed method is superior to the baseline in both the word ordering and sentence ordering tasks.

Task	Method	Error rate	Avg. Hamming distance	Avg. Kendall distance
Word ordering	Baseline	0.350	0.800	0.600
	Ours	0.200	0.500	0.350
Sentence ordering	Baseline	0.769	2.231	1.692
	Ours	0.615	1.923	1.385

skills of individual workers. Moreover, a number of malicious workers participate in crowdsourcing [15]. They are motivated by financial rewards and try to complete the tasks as quickly as possible with the minimum effort by providing illogical submissions.

A widely used approach is to obtain multiple submissions from different workers and aggregate them by applying a majority vote [5] or other rules. Dawid and Skene addressed the problem of aggregating the medical diagnoses of multiple doctors to achieve more accurate decisions [6]. Smyth et al. applied the method to the problem of inferring the true labels of images from multiple noisy labels [16]. Whitehill et al. explicitly modeled the difficulty of each task [7], and Welinder et al. introduced the difficulty of each task for each worker [8]. The usage of these methods is limited to the tasks that constitute binary or multiple-choice questions; however, the tasks in crowdsourcing comprise varied types of questions. A few methods have been proposed to extend the applicability of the aggregation-based quality control approach to more general crowdsourcing tasks [9].

Although the probabilistic models for ranking have been widely studied [11], only a few studies in the literature focused on item ordering tasks in the context of crowdsourcing. Chen et al. proposed a quality control method for item ordering tasks [13] based on a pairwise ranking model; however, their method aims to find the correct ordering of a single, large set of items, while our method focuses on solving multiple different (relatively small) ordering tasks simultaneously. Additionally, since their method is based on pairwise comparisons, it is not always suitable for tasks where more than two items are needed to decide their correct positions. Wu et al. also employed the general distance-based model in the context of *learning to rank* from multiple annotators [17], while our approach employs a more specific distance measure, i.e., Spearman distance, so that the inference is more simple and efficient.



(a) Word ordering (b) Sentence ordering

Fig. 5. Accuracy evaluation of estimated worker abilities. Relations between averaged squared error of each worker's responses to ground truths and estimated worker ability are shown. Strong negative correlations (-0.853 and -0.695 for the word ordering tasks and the sentence ordering tasks, respectively) are confirmed.

7 Conclusion

We addressed the problem of quality control for item ordering tasks in crowdsourcing, where multiple workers are asked to perform each task, which consists of positioning given items in the correct order. By extending a distance-based probabilistic ordering model to incorporate the ability of each worker, we built our proposed method for aggregating the collected orders to obtain more accurate orders in a setting where variability in the workers' abilities exists. We also introduced an efficient algorithm to estimate the true orders that employs the Spearman distance as the distance measure in a distance-based ordering model. Experimental results on two kinds of crowdsourcing tasks, word ordering tasks and sentence ordering tasks, showed that our method successfully achieved more accurate orders than the baseline method, which does not consider the worker's ability.

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

References

1. Bernstein, M., Little, G., Miller, R., Hartmann, B., Ackerman, M., Karger, D., Crowell, D., Panovich, K.: Soylent: A word processor with a crowd inside. In: Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, UIST (2010)
2. Bigham, J., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R., Miller, R., Tatarowicz, A., White, B., White, S., 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. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: 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)
4. Sorokin, A., Forsyth, D.: Utility data annotation with Amazon Mechanical Turk. In: Proceedings of the 1st IEEE Workshop on Internet Vision (2008)
5. Sheng, V.S., Provost, F., Ipeirotis, P.G.: 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)
6. Dawid, A.P., Skene, A.M.: 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)
7. Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., Movellan, J.: Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In: Advances in Neural Information Processing Systems, vol. 22 (2009)
8. Welinder, P., Branson, S., Belongie, S., Perona, P.: The multidimensional wisdom of crowds. In: Advances in Neural Information Processing Systems, vol. 23 (2010)
9. Lin, C., Mausam, M., Weld, D.: Crowdsourcing control: Moving beyond multiple choice. In: Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence, UAI (2012)
10. Zhang, H., Law, E., Miller, R., Gajos, K., Parkes, D., Horvitz, E.: Human computation tasks with global constraints. In: Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems, CHI (2012)
11. Marden, J.I.: Analyzing and Modeling Rank Data, vol. 64. CRC Press (1995)
12. Mallows, C.L.: Non-null ranking models. I. Biometrika 44, 114–130 (1957)
13. Chen, X., Bennett, P.N., Collins-Thompson, K., Horvitz, E.: Pairwise ranking aggregation in a crowdsourced setting. In: Proceedings of the 6th ACM International Conference on Web Search and Data Mining, WSDM (2013)
14. Chang, P.C., Toutanova, K.: A discriminative syntactic word order model for machine translation. In: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL (2007)
15. Eickhoff, C., de Vries, A.: How crowdsourcable is your task? In: Proceedings of the Workshop on Crowdsourcing for Search and Data Mining, CSDM (2011)
16. Smyth, P., Fayyad, U., Burl, M., Perona, P., Baldi, P.: Inferring ground truth from subjective labelling of venus images. In: Advances in Neural Information Processing Systems, vol. 7 (1995)
17. Wu, O., Hu, W., Gao, J.: Learning to rank under multiple annotators. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI), pp. 1571–1576 (2011)