Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6590–6596
Marseille, 20-25 June 2022

Quality Control for Crowdsourced Bilingual Dictionary in Low-Resource Languages

Hiroki Chida, Yohei Murakami, Mondheera Pituxcoosuvarn
Ritsumeikan University
1–1–1 Noji-Higashi, Kusatsu, Shiga, 525–8577 Japan
is0363xp@ed.ritsumei.ac.jp, yohei@fc.ritsumei.ac.jp, mondheera@fc.ritsumei.ac.jp

Abstract

In conventional bilingual dictionary creation by using crowdsourcing, the main method is to ask multiple workers to translate the same words or sentences and take a majority vote. However, when this method is applied to the creation of bilingual dictionaries for low-resource languages with few speakers, many low-quality workers are expected to participate in the majority voting, which makes it difficult to maintain the quality of the evaluation by the majority voting. Therefore, we apply an effective aggregation method using a hyper question, which is a set of single questions, for quality control. Furthermore, to select high-quality workers, we design a task-allocation method based on the reliability of workers which is evaluated by their work results.

Keywords: crowdsourcing, bilingual dictionary, quality control, task assignment

1. Introduction

Recently, crowdsourcing is becoming mainstream to create language resources including bilingual dictionaries. Crowdsourcing is a scheme for requesting work from a large and open group of people via the internet, and it can be used to order a large number of works that require human labor. Crowdsourcing is especially used to request relatively difficult tasks for computers, but not so difficult for humans. However, in crowdsourcing, where the tasks are executed by an unspecified number of workers the abilities of whom vary, it is difficult to guarantee the quality of the execution results. Especially in the case of bilingual dictionaries creation between low-resource languages (Murakami, 2019), the number of people who can speak multiple low-resource languages is limited, and the average ability of workers is low. This results in the method of assigning the same task to multiple workers and using majority voting has a high possibility of obtaining wrong answers, and quality control cannot be performed well.

Therefore, we aim to improve quality in an environment with a small number of highly reliable workers by using an answer aggregation method on hyper questions (multiple tasks considered together as one task). Since workers with high ability tend to agree on the answers to hyper questions, the method increases the possibility that workers with high ability will be in the majority. To this end, we address the following two problems.

Selecting highly reliable evaluators

In the answer aggregation method on hyper questions, it is assumed that a small number of high-quality workers are involved. Therefore, it is necessary to select highly reliable evaluators from a crowd.

Reducing the number of tasks

Even if a worker is able to correctly evaluate whether a bilingual text is correct or not, in the case of wrong bilingual texts, the worker may have to redo the translation, which increases the number of tasks.

For these problems, we dynamically evaluate the reliability of workers based on their work results, and selected workers who were estimated to be highly skilled. Specifically, we set a parameter ‘Reliability’ for each worker and increased or decreased the reliability based on the task results. In addition, we adjust the probability of task assignment based on the reliability of each worker.

2. Related Work

2.1. Language Resource Creation using Crowdsourcing

The mainstream method of creating language resources is to ask experts to do so, and this is known to be costly. However, the use of crowdsourcing has made it possible to create language resources at a relatively low cost, and a variety of research on language resources has been conducted. For example, a method for creating bilingual examples between English and Spanish using Amazon Mechanical Turk [1] (AMT) has been proposed (Negri and Mehdad, 2010).

In the low-resource language domain, there is an approach to create a bilingual dictionary A-C automatically with bilingual dictionaries A-B and B-C as inputs (Nasution et al., 2017). However, it is difficult to complete a bilingual dictionary with only machines, so they also combined it with crowdsourcing (Nasution et al., 2021).

[1] Amazon Mechanical Turk (https://www.mturk.com)

2.2. Quality Control on Crowdsourcing

One of the most important research topics in crowdsourcing is quality control. Since tasks are performed by humans, it is not always possible to obtain correct results. In addition, since tasks are requested from an unspecified number of people, there is a possibility that workers with low ability or workers who intentionally perform low-quality work (spammers) will perform tasks. Therefore, the quality of the results cannot be guaranteed only by the results of a single worker. In the research of quality control, there are two main approaches: an approach to aggregate work results for improving the overall quality and an approach to improve the quality of individual work results.

The former is mainly an approach that attempts to obtain high-quality results by removing errors from the work results. As an example, the method of assigning the same task to multiple workers, and then taking a majority vote is used. However, the majority voting can lead to the correct answer when the ability of the workers is high, while it is difficult to obtain the correct answer when the ability of the workers is low (less than 50% correct in the case of binary choice type tasks) (Sheng et al., 2008). For such cases where experts are in the minority, an answer aggregation method using hyper questions has been proposed as an effective method (Li et al., 2017). A hyper question is a set of single questions, in which multiple questions are considered together as one. Since experts are more likely to agree on the answers to multiple questions than non experts, majority voting on hyper questions is particularly effective when there are few workers with high ability.

The latter is an approach that attempts to improve the results of task execution itself by designing rewards and tasks or selecting workers before requesting workers to perform tasks. Especially, the method of extracting workers who are estimated to have high ability in advance and assigning tasks to the extracted workers is expected to improve the quality of the work results, because it can eliminate low ability workers and spammers before executing the task, and only workers who are estimated to have a high ability can actually perform the task.

2.3. Task Assignment

In the task assignment, it is necessary to estimate the abilities of workers in advance in order to extract workers who can be expected to deliver high-quality work results. However, it is difficult to know the abilities of workers in advance because the abilities of workers in crowdsourcing vary widely.

Therefore, a method of detecting workers with high ability by using a task the correct answer of which is known in advance (gold task) has been used. For example, there are two methods: one is to assign gold tasks in advance and filter workers by evaluating their answers, and the other is to blend gold tasks into normal tasks to measure and select the ability of workers (Kazai et al., 2011). When a worker is judged to have a low ability by these methods, it is possible to take countermeasures such as not assigning tasks to the worker afterward, placing restrictions on some tasks, or not using the results of the worker's output. These methods are considered to be the most effective ways of estimating the abilities of workers when the average ability of workers is not high. However, if the gold tasks are mixed in with the actual tasks, the reward for answering the gold tasks, whose answers are already known, must be paid, which reduces the cost-effectiveness of the method. In the case of measuring workers' abilities in advance, it is necessary to assign gold tasks to all workers, which simply reduces the efficiency of the workload. Furthermore, it is known that it is very difficult and costly to generate gold tasks, so a method to automatically generate gold tasks based on data collected has been proposed (Oleson et al., 2011)

In this paper, we assume the bilingual dictionaries creation using crowdsourcing in low-resource languages. Therefore, the number of workers who can speak these languages is small, and the average ability of workers is not high. Therefore, we aim to improve the quality of the created bilingual dictionary by combining an answer aggregation method that is effective even for such a crowd with low average ability and a task assignment method based on workers' reliability calculated from the results of each worker's work.

3. Workflow for Bilingual Dictionary Creation

3.1. Workflow

Considering a workflow consisting of a translation task and multiple evaluation tasks (Figure 1), we ensure redundancy by performing multiple evaluation tasks for each bilingual creation task. In other words, the final evaluation of the bilingual text produced by a translation task is determined by a majority vote of the results of evaluation tasks. If a 'Correct' bilingual text is produced and it is evaluated as 'Correct', the 'Correct' bilingual text is obtained. If a 'Wrong' bilingual text is produced and it is evaluated as 'Wrong', the 'Wrong' translation is obtained. Otherwise, the bilingual text is not acquired. If no bilingual text is obtained, the process is repeated from a translation task until bilingual texts for all words are obtained.

3.2. Tasks

We assume that there are two types of tasks assigned to workers: a translation task, which is a free input task to create a bilingual text from a given word or sentence, and an evaluation task, which is a binary-choice task to evaluate whether the bilingual text created by a translation task is 'Correct' or 'Wrong.'

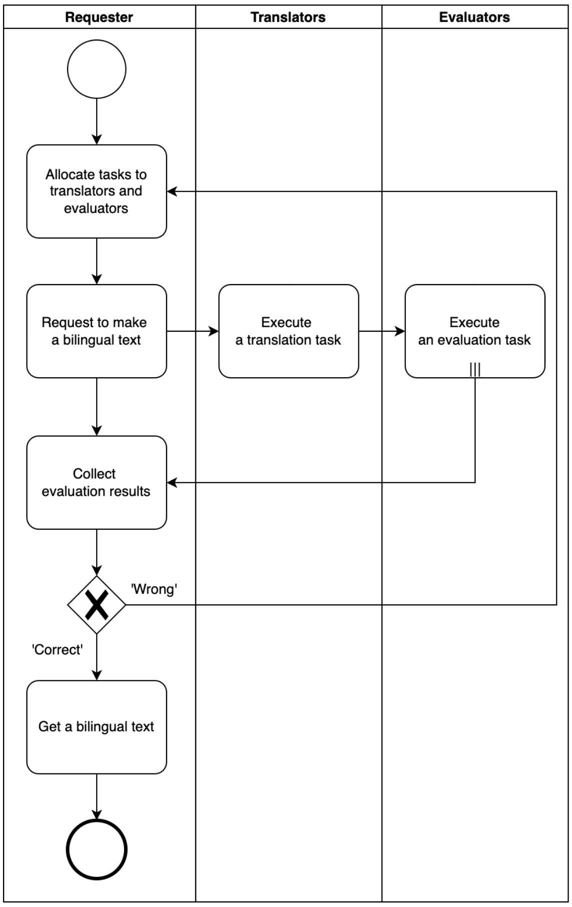

Requester
Translators
Evaluators
Allocate tasks to translators and evaluators
Request to make a bilingual text
Execute a translation task
Execute an evaluation task
Collect evaluation results
'Wrong'
'Correct'
Get a bilingual text

Figure 1: Workflow for bilingual dictionary creation

4. Answer Aggregation on Hyper Questions

4.1. Hyper Questions

The common aggregation methods such as majority voting often fail when the majority of workers do not know the correct answers. To emphasize the answers of a few high-quality workers, The aggregation methods on hyper questions are proposed (Li et al., 2017).

A hyper question consists of a subset of original single questions, and an answer to a hyper question is a set of answers to the questions included in the hyper question. A set of k original single questions is defined as a k-hyper question

4.2. Majority Voting on Hyper Questions

As the specific answer aggregation method on Hyper questions, we use majority voting on hyper questions for evaluation tasks.

At first, a set of some evaluation tasks Q is created, and k-hyper questions are constructed by combining single evaluation tasks in Q Then taking majority voting to each hyper question, which results in an answer to the hyper question. The aggregated results of the hyper questions are decoded into answers to the single questions. Finally, another round of majority voting is carried out for each question. Consequently, the results of the first round of majority voting on hyper questions are aggregated to obtain the final answer for every single question.

Figure 2 shows the procedure of majority voting on hyper questions, which consists of five evaluators and four evaluation tasks in which the evaluators determine whether each bilingual text (Japanese - English) is ‘〇（correct）’ or ‘×（wrong）’. In this example, k is set to 3. ‘〇’ is the correct answer for all of the evaluation tasks. In the first step, four 3-hyper questions are created from the four evaluation tasks. An answer to a hyper question is the concatenation of the answers to the constituent single evaluation task. In the second step, taking majority voting to each hyper question; in this case, the answer ‘〇〇〇’ is chosen for all the hyper questions. In the third step, each of the majority answers to the hyper questions votes for the single evaluation task included in it. Finally, in the fourth step, another round of majority voting aggregates the votes to the single evaluation task to obtain the final answers. Simple majority voting fails in the evaluation task for bilingual text2, but majority voting on hyper questions succeeds. If there are no majority answers in the second step and some of the single evaluation tasks do not get the answers, another round of majority voting is taken by the evaluators who voted majority answers for the rest of succeeded evaluation tasks.

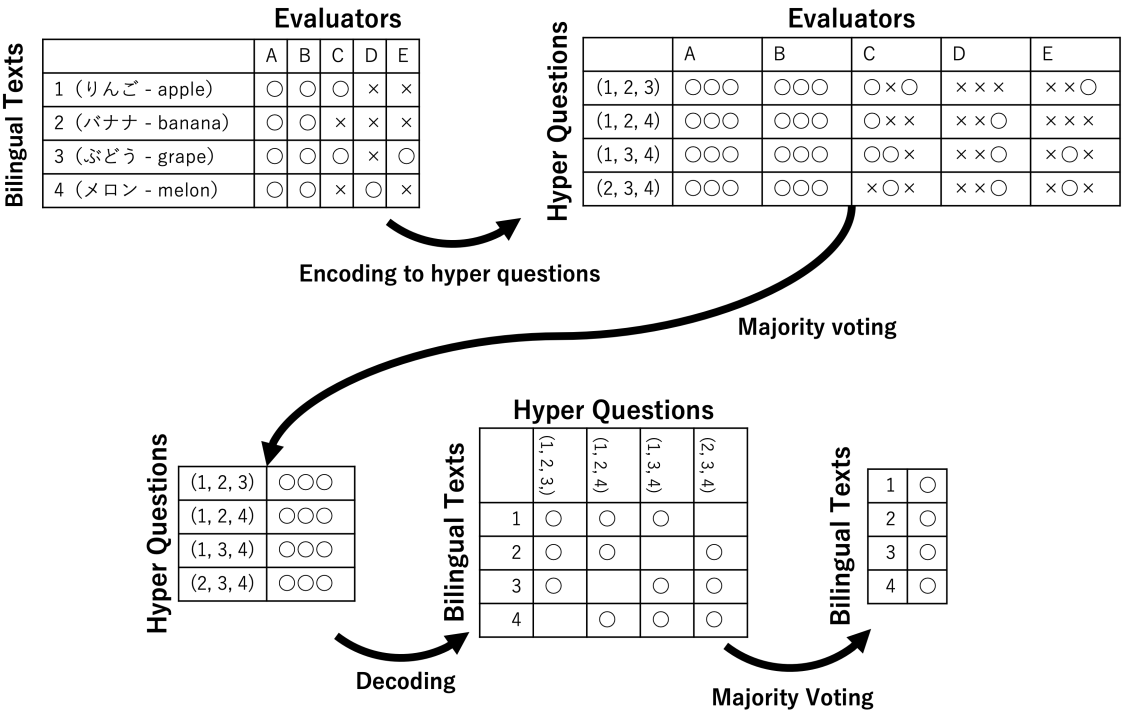
5. Task Assignment Based on Workers' Reliability

In this research, we aim to improve the quality and reduce the cost of crowdsourcing by identifying workers who are estimated to be highly skilled based on their work results and proactively assigning tasks to them. For this purpose, we propose a method to dynamically evaluate the reliability of workers based on their work results.

5.1. Workers' Reliability

A parameter 'Reliability' is set for each worker, and the initial value is 0. Reliability is calculated based on the results of translation tasks and evaluation tasks as follows.

- If the bilingual text created by a translation task is evaluated as 'correct' by evaluation tasks, the reliability of the translator is increased by $+1$.
- If the bilingual text created by a translation task is evaluated as 'wrong' by evaluation tasks, the reliability of the translator is increased by -1.
- If a worker's evaluation of all the created bilingual texts in a given task set Q is a majority of the final evaluation obtained from the aggregation of the evaluation tasks, the reliability of the evaluator is increased by $+1$.
- If a worker's evaluation of all the created bilingual texts in a given task set Q is a minority of the



	A	B	C	D	E
1（りんご - apple)	○	○	○	×	×
2（バナナ - banana)	○	○	×	×	×
3（ぶどう - grape)	○	○	○	×	○
4（メロン - melon)	○	○	×	○	×

	A	B	C	D	E
(1, 2, 3)	○○○	○○○	○×○	×××	××○
(1, 2, 4)	○○○	○○○	○××	××○	×××
(1, 3, 4)	○○○	○○○	○○×	××○	×○×
(2, 3, 4)	○○○	○○○	×○×	××○	×○×

(1, 2, 3)	○○○
(1, 2, 4)	○○○
(1, 3, 4)	○○○
(2, 3, 4)	○○○

	(1, 2, 3,)	(1, 2, 4)	(1, 3, 4)	(2, 3, 4)
1	○	○	○	
2	○	○		○
3	○		○	○
4		○	○	○

1	○
2	○
3	○
4	○

Figure 2: Example of majority voting on hyper questions procedure

final evaluation obtained from the aggregation of the evaluation tasks, the reliability of the evaluator is increased by -1.

This calculation is performed each time the evaluation of all the created translations in one problem set Q is completed.

5.2. Task Assignment

By using the Reliability of each worker, we proposed two types of task assignment methods

- Assigning evaluation task using threshold
- Task assignment using weighted probabilities

For the first method, we placed restrictions on workers to allocate evaluation tasks. For the bilingual evaluation task, we consider a worker whose reliability is 1 or higher to be a trusted worker, and only trusted workers can perform evaluation tasks. This method is expected to reduce the number of errors in evaluation tasks.

For the second method, the probability of task assignment for both translation tasks and evaluation tasks is adjusted based on the weight of each worker using his/her reliability. When the total number of workers who can perform a task is n, the weight w_i of the ith worker is calculated as in Equation (1).

$$w_i = 1 + r_i - r_{min} \tag{1}$$

The r_i shows the reliability of the ith worker, and the r_{min} is the lowest reliability among all workers who can perform the task. By calculating the weight as in Equation (1), we can avoid that the weight of the worker with the lowest reliability becomes 0 (the probability of being assigned the task becomes 0). As the work progresses, the difference in the weights increases as the difference in the reliability among the workers becomes larger.

The probability that a task is assigned to a worker, p_i, can be calculated by using weights, as in Equation (2).

$$p_i = \frac{w_i}{w_1 + w_2 + w_3 + \cdots + w_i + \cdots + w_n} \tag{2}$$

By performing these calculations each time a task is assigned, we can make it easier to assign a task to a worker with high reliability (a worker who is estimated to be highly capable) and harder to assign a task to a worker with low reliability (a worker who is estimated to be less capable), thereby automatically eliminating workers who are estimated to be less capable. This can be expected to improve accuracy and reduce costs.

6. Evaluation

6.1. Modeling

For the evaluation, we modeled crowdsourcing workers and tasks assuming creating a bilingual dictionary for a low-resource language.

6.1.1. Workers

The higher ability of the worker, the quality of the task execution result is higher. In this paper, the ability of a worker is defined as the vocabulary in multiple languages and is represented by $x(0 \leq x \leq 1)$. When x

is closer to 1, the worker recognizes more vocabulary, and the more likely he/she is to perform the task correctly. On the other hand, when x is closer to 0, the worker recognizes less vocabulary, and the possibility that the task will be incorrect increases. For simplicity, we assume that the quality of the task execution result is probabilistically determined by the ability of a worker. In this paper, we follow previous studies and represent the ability of a worker using a beta distribution. The probability density function $f(x|a, v)$ is represented by equation 3 (Goto et al., 2016).

$$f(x|a, v) = \mathrm{Beta}\left(\frac{a}{\min(a, 1-a)v}, \frac{1-a}{\min(a, 1-a)v}\right) \quad (3)$$

$a \in (0, 1)$ is the normalized value of workers ’ ability, and $v \in (0, 1)$ is the parameter that determines the variance of workers ’ ability. When v is closer to 0, the variance is closer to 0, and when v is closer to 1, the variance in the beta distribution with the average a is larger. The above model of workers was adopted by (Goto et al., 2016).

6.1.2. Tasks

We assume that the result of a translation task is ‘Correct’ if the worker knows the translation of the given word, and ‘Wrong’ if the worker does not know the translation of the given word. Therefore, it is completely dependent on the ability of the worker whether a correct bilingual text is produced or not (Figure 3). However, since an evaluation task is a binary choice task, if the worker knows the correct translation for a given word, he/she will evaluate it as ‘Correct’. However, if the worker does not know the translation of the word, he/she will randomly select one of the two values ‘Correct’ or ‘Wrong’ (Figure 4). Therefore, in an evaluation task, no matter how low the ability of the worker is, it is guaranteed that the worker will make a ‘Correct’ evaluation with a probability of more than 50%.

6.2. Evaluation Method

The methods, including the proposed method, are evaluated in terms of the accuracy of the produced bilingual texts and the work quantity required to obtain all the bilingual texts.

1. Proposed Method (Reliable_hyper)
   A model that combines the answer aggregation on Hyper questions and the task assignment based on workers’ reliability

2. Comparison Method1 (Random_hyper)
   A model using the answer aggregation on Hyper questions
   All tasks are assigned randomly from the entire workers

3. Comparison Method2 (Reliable)
   A model using the task assignment based on workers’ reliability
   A simple majority voting is used to aggregate the results of evaluation tasks

4. Comparison Method3 (Random)
   A model that randomly assigns tasks from the entire workers and uses simple majority voting to aggregate the results of evaluation tasks

In order to measure the performance of each method described above, we use the following indicators.

1. Accuracy of the produced bilingual texts
   The accuracy of the produced bilingual texts by each method is calculated as follows.

   $$\text{Accuracy} = \frac{\text{Number of bilingual texts produced correctly}}{\text{Total number of obtained bilingual texts}} \quad (4)$$

   This indicator helps to compare the simple quality of the outputs from each method.

2. Work quantity required to obtain all the bilingual texts.
   The work quantity is the total unit times of the translation tasks and the evaluation tasks which are executed until all the bilingual texts are obtained. A unit time is calculated from the estimated time taken for doing the task. Since translation tasks are more difficult than evaluation tasks, we defined that a translation task takes 3 units and an evaluation task takes 1 unit. The cost model was adopted by (Nasution et al., 2021). This indicator helps to compare the efficiency and cost of each method.

In order to evaluate the indicators described above, we conducted simulations using each method. We set the number of workers to 20 and assumed that there are 1000 target words. The ability of each worker is determined based on the model in 6.1.1, and we varied the average of workers’ abilities between 0.2 and 0.7 with the variance 0.5. To eliminate bias due to random numbers, we used the average of the results of 100 simulations for each method.

6.3. Results

6.3.1. Accuracy

The accuracy of the proposed method, Reliable_hyper, was the highest, followed by Reliable, Random_hyper, and Random. The difference between Reliable_hyper, which had the highest accuracy, and Reliable, the second highest, was about 5-10%, as illustrated in Figure 5.

6.3.2. Work Quantity

The work quantity tended to be larger for Reliable_hyper and Random_hyper, which are the models using the answer aggregation on Hyper questions. However, for Reliable_hyper, the work quantity was the smallest when the average of the workers’ ability was 0.5 or higher, as shown in Figure 6.

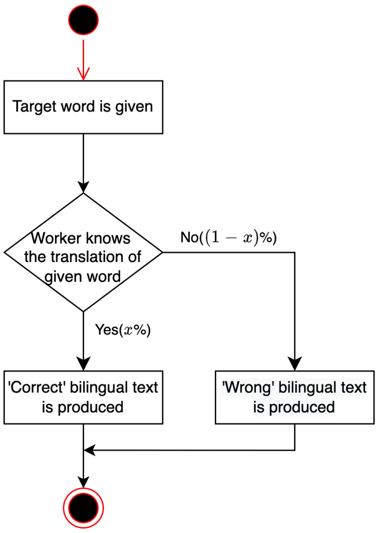

Target word is given
Worker knows the translation of given word
No($(1-x)$%)
Yes(x%)
'Correct' bilingual text is produced
'Wrong' bilingual text is produced

Figure 3: A translation task model

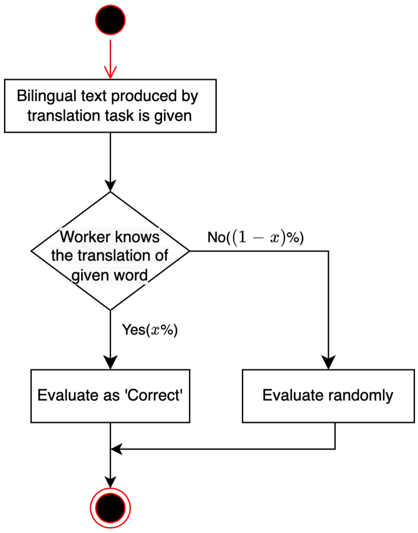

Bilingual text produced by translation task is given
Worker knows the translation of given word
No($(1-x)$%)
Yes(x%)
Evaluate as 'Correct'
Evaluate randomly

Figure 4: An evaluation task model

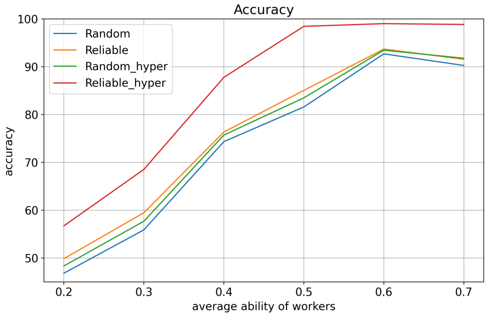


Figure 5: Accuracy

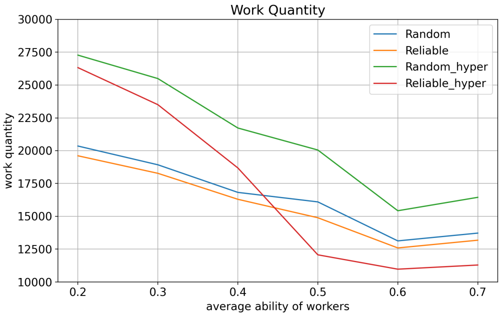


Figure 6: Work quantity

6.4. Discussion

6.4.1. Accuracy

Both Reliable and Random_hyper were more accurate than Random, indicating that the task assignment based on workers' reliability and the answer aggregation on hyper questions are effective. In addition, when we compared Reliable and Random_hyper, the accuracy of Reliable was higher than that of Random, indicating that it is more effective to assign tasks to workers with high ability than to improve the quality of answer aggregation. Furthermore, the accuracy of Reliable_hyper, which combines the task assignment based on workers' reliability and the answer aggregation on hyper questions, was particularly high, indicating that these methods are more effective when combined than when used individually.

6.4.2. Work Quantity

Since the work quantity for Reliable_hyper and Random_hyper, which use the answer aggregation on hyper questions, tended to be larger, it is easy to assume that many redos occurred. This may be because the majority voting on hyper questions makes it more difficult to aggregate the answers than in simple majority voting. Therefore the integrations of evaluation tasks often fail. However, when the average of workers' ability was 0.5 or higher, the work quantity for Reliable_hyper was the smallest. This shows that if evaluation tasks can be assigned to high-quality workers from a crowd with more than a certain number of high ability workers, the majority voting on hyper questions is more likely to be successful, and redoing the task is less likely to occur. Furthermore, in Reliable_hyper, translation tasks are also assigned preferentially to the worker with the highest reliability, so there are few wrong bilingual texts created in the first place. Regarding the number of reliable workers, whose abilities are more than 0.7, there were two reliable workers when the average of workers' abilities was 0.4, and there were four reliable workers when the average of workers' abilities was 0.5, as illustrated in Figure 7 and 8. This shows that two reliable workers are too few because they are mostly assigned to translation tasks and not to evaluation tasks so that the majority voting on hyper questions does not work well even if they execute translation tasks very well.

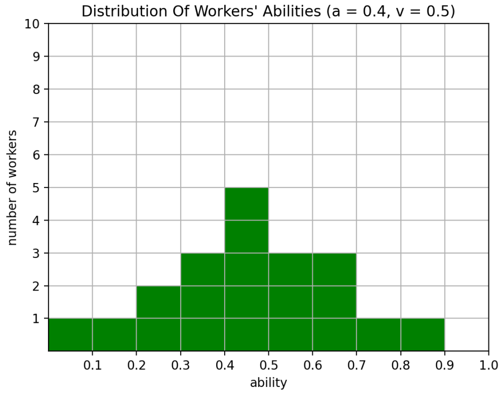


Figure 7: Distribution of workers' abilities ($a = 0.4, v = 0.5$)

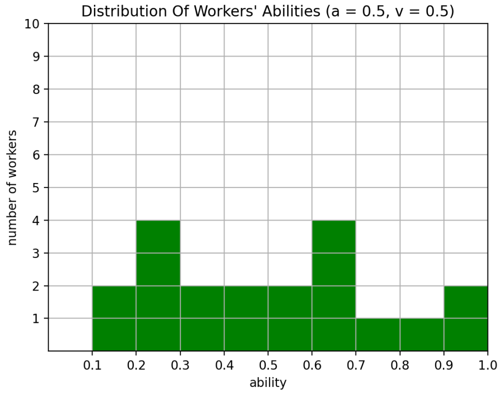


Figure 8: Distribution of workers' abilities ($a = 0.5, v = 0.5$)

7. Conclusions

In this research, we showed that higher quality bilingual texts can be obtained by combining a task assignment method based on the reliability of workers, which is dynamically evaluated by their work results, and an answer aggregation method on hyper questions. This is the result of improving the quality of integrating the answers of evaluation tasks by majority voting on hyper questions and assigning tasks to workers estimated to have high ability based on their reliability. In addition, we succeeded in reducing the work quantity due to redoing by proactively assigning translation tasks to workers who were estimated to have the high ability. As a result, the proposed method, which combines a task assignment method based on workers' reliability and an answer aggregation method on a hyper question, succeeded in obtaining 5-10% higher accuracy than the case when these methods are used individually, while reducing the work quantity due to the use of majority voting on hyper questions.

Acknowledgments

This research was partially supported by a Grant-in-Aid for Scientific Research (B) (21H03561,2021-2024, 21H03556, 2021-2023) and a Grant-in-Aid Young Scientists(A)(17H04706,2017-2020) from the Japan Society for the Promotion of Science(JSPS).

References

Goto, S., Ishida, T., and Lin, D. (2016). Understanding crowdsourcing workflow: modeling and optimizing iterative and parallel processes. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*.

Kazai, G., Kamps, J., Koolen, M., and Milic-Frayling, N. (2011). Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 205–214.

Li, J., Baba, Y., and Kashima, H. (2017). Hyper questions: Unsupervised targeting of a few experts in crowdsourcing. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1069–1078.

Murakami, Y. (2019). Indonesia language sphere: An ecosystem for dictionary development for low-resource languages. In *Journal of Physics: Conference Series*, volume 1192, page 012001. IOP Publishing.

Nasution, A. H., Murakami, Y., and Ishida, T. (2017). A generalized constraint approach to bilingual dictionary induction for low-resource language families. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(2):1–29.

Nasution, A. H., Murakami, Y., and Ishida, T. (2021). Plan optimization to bilingual dictionary induction for low-resource language families. *Transactions on Asian and Low-Resource Language Information Processing*, 20(2):1–28.

Negri, M. and Mehdad, Y. (2010). Creating a bilingual entailment corpus through translations with mechanical turk: $100 for a 10-day rush. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon' s Mechanical Turk*, pages 212–216.

Oleson, D., Sorokin, A., Laughlin, G., Hester, V., Le, J., and Biewald, L. (2011). Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 43–48.

Sheng, V., Provost, F., and Ipeirotis, P. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622, 08.