Crowdsourcing for Evaluating Machine Translation Quality

Shinsuke Goto, Donghui Lin, Toru Ishida
Department of Social Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-Ku, Kyoto 6068501, Japan
s-goto@ai.soc.i.kyoto-u.ac.jp, {lindh, ishida}@i.kyoto-u.ac.jp

Abstract

The recent popularity of machine translation has increased the demand for the evaluation of translations. However, the traditional evaluation approach, manual checking by a bilingual professional, is too expensive and too slow. In this study, we confirm the feasibility of crowdsourcing by analyzing the accuracy of crowdsourcing translation evaluations. We compare crowdsourcing scores to professional scores with regard to three metrics: translation-score, sentence-score, and system-score. A Chinese to English translation evaluation task was designed using around the NTCIR-9 PATENT parallel corpus with the goal being 5-range evaluations of adequacy and fluency. The experiment shows that the average score of crowdsource workers well matches professional evaluation results. The system-score comparison strongly indicates that crowdsourcing can be used to find the best translation system given the input of 10 source sentence.

Keywords: Evaluation, Crowdsourcing, Machine Translation

1. Introduction

The demand for translation evaluations has been growing due to the increasing number of machine translation systems. While many methods have been developed to evaluate translation quality, the traditional technique remains manual evaluations by bilingual professionals.

Quality assessment has several objectives. In many cases, evaluation score of each translation is used to indicate the quality of translation (Banerjee and Lavie, 2005). Shi et al. uses an automatic evaluation method to select the best translation among those generated by multiple machine translators (Shi et al., 2012). Goto et al. conducted an evaluation task of 5-range evaluation to compare the quality of translation systems by using the average score on each source sentence (Goto et al., 2011).

In addition, some studies have focused on the automatic evaluation of translations (Papineni et al., 2002) (Banerjee and Lavie, 2005). Their proposals offer good correlation with professionals, but require multiple reference translations to work well. Several studies have pointed out the limitations of automatic translation evaluation methods (Zhang et al., 2004) (Callison-Burch et al., 2006).

In this study, we propose to use crowdsourcing to achieve low-cost evaluations with the same quality as professional 5-level evaluations. Crowdsourcing enables employers to allocate tasks to anonymous workers on the Web. Unlike professionals, crowdsourcing can yield quick and low-cost evaluations. A good example of crowdsourcing-based evaluation was provided by Chen et al., who used crowdsourcing to measure the QoE (Quality of Experience) of networks (Chen et al., 2010). If these approaches are applicable to the translation domain, we can create easy and low-cost methods for creating evaluation sets of machine translation or machine translation corpora.

However, there are no criteria for evaluating non-expert evaluations since manual evaluation is considered as the standard for evaluation. Translation evaluation is used in several ways, but it remains unclear how to use crowdsourcing and what kind of evaluations are possible. In addition, the quality of workers in crowdsourcing is generally not guaranteed. Tasks in crowdsourcing do not require workers of special ability, and the resulting quality is assured only by redundancy in most cases. Redundancy is achieved by majority voting, so more than half the workers need to be accurate. An analysis of the applicable fields of translation is also required. Of particular note, the crowdsourcing market is bedeviled by spammers, who aim to gain illicit rewards. Existing works of crowdsourcing translation evaluation require reference translations made by experts, and does not expect workers to have professional ability (Callison-Burch, 2009) (Bentivogli et al., 2011).

Based on these difficulties of crowdsourcing, we address the two issues to utilize crowdsourced evaluations for translation quality assessment. The first is the Comparison scheme to clarify the usefulness of crowdsourcing evaluations. We have to create a feasible way of using crowdsourcing evaluations of translations. Therefore, a scheme that compares crowdsource workers to professionals is required. The second issue is analysis of crowdsourcing evaluations and professional evaluations. Up to now, crowdsourcing evaluation is not compared to the professionals in the tasks which require expertise. We have to compare the result of crowdsourcing and professional evaluation. In addition, we have to determine the low-cost evaluation of crowdsourcing can replace the professional evaluation.

We analyze the translation evaluations created by crowdsourcing and determine whether crowdsourcing can substitute for professional evaluations. Based on professional evaluations, three performance metrics are employed: translation-score, sentence-score, and system-score comparison, which indicate the closeness of the absolute evaluation score, the fitness of the relative order in one source sentence, and ranking by translation systems, respectively.

This paper is structured as follows: Section 2 introduces related works on translation evaluation methods and crowdsourcing. Next, a scheme to compare crowdsourcing to expert evaluations will be explained in Section 3, then Section 4 shows the experimental settings. The results will be analyzed on Section 5 and discussed on Section 6. Finally, Section 7 concludes this paper.

2. Related Works

This section shows related works on crowdsourcing evaluations. They include translation evaluations, crowdsource evaluations, and crowdsource translations. Up to now, translation evaluations are performed by experts. Adequacy and fluency are the general criteria for manual evaluations (White et al., 1994). Also, there exist several automatic evaluation methods. Popular automatic evaluations are NIST, BLEU, and METEOR (Doddington, 2002) (Banerjee and Lavie, 2005) (Papineni et al., 2002). These methods are based on forming N-grams between each machine-translated sentence and one or more reference translations. On the other hand, some studies address the problems of these methods. For example, Zhang, et al. pointed out that BLEU and NIST have different ranking schemes (Zhang et al., 2004). Also, Callison-Burch et al. described how the assessment quality of BLEU depends on the multiple reference translations (Callison-Burch et al., 2006). They argued the translations of a sentence generated by different machine translators can be definitely different by human, especially in case of long sentences.

Crowdsourcing is also used in various evaluations. For example, Chen, et al. has measured the quality of the experience offered by crowdsource workers (Chen et al., 2010). Workers were assigned to the task of pairwise comparison of six movies, and the proposed system standardized the comparison results. In addition, Gabriella et al. evaluated book search relevance as determined by crowdsourcing (Kazai et al., 2011) Workers participated in several experiments on the relationship between search query and search result. This paper concluded that both the qualification before participating and the design of task should be considered carefully. Voting by crowdsourcing can be also regarded as evaluation. For example, voting is sometimes performed during complex workflows performed by the crowdsourcing (Bernstein et al., 2010) (Little et al., 2010).

Crowdsourcing studies have also addressed translation. Zaidan et al. showed the feasibility of crowdsourced translation using a sequence of tasks (Zaidan and Callison-Burch, 2011). Workers create translation drafts, edit the translated sentences, and then vote to select the best translation. Ambati et al. proposed a combined of active learning and crowdsourced translation in order to improve the quality of statistical machine translation (Ambati et al., 2010). In addition, Aziz, et al. developed and investigated a crowdsourcing-based tool that enables post-editing of machine translations and evaluation of machine translations (Ambati et al., 2010). In a discussion of nonprofessional cost, Lin et al. suggest that low-cost language resource creation can be easily realized by utilizing non-experts (Lin et al., 2010). Morita et al. introduced a protocol to create translations by a machine translation system and two monolingual non-experts (Morita and Ishida, 2009). Green et al. analyzed the effects of post-editing (Green et al., 2013). Examples of previous works on the application of crowdsourcing to translation evaluation include Callison-Burch et al. and Luisa et al. (Callison-Burch, 2009) (Bentivogli et al., 2011). Callison-Burch et al. designed two evaluation tasks around professional reference translations. Experiments showed that crowdsourcing evaluations are useful in comparing translation systems. Luisa et al. designed a ranking task for machine translation systems. They showed that Spearman ’ s rank correlation is better than automatic evaluation systems like BLEU or NIST. These studies assumed the existence of reference translations and that monolingual workers were performing the tasks.

This research addresses translation evaluation without reference sentences. Another goal is to obtain absolute evaluation scores for each translation. Therefore, crowdsource workers are assigned the same evaluation task as professionals, and we compare and analyze the crowdsourced and professional evaluations.

3. Evaluation Scores to Compare

To compare the evaluations between crowdsourcing and professional, we need to define the metrics to determine the usage of crowdsourcing evlauations. We use three techniques for comparing crowdsourced and professional evaluations. In this paper, we focus on the determination of the effective fields of crowdsource evaluations. The comparisons indicate the closeness of absolute evaluation scores, the closeness of the ranking of machine translations in the source sentence, and the closeness of the ranking of machine translations in the given set of translations. The evaluation scores are called translation-score, sentence-score, and system-score, respectively.

We compare the scores made by crowdsource workers and a professional. Translation-score is the absolute score of translation for each translation, and is used in the evaluation of machine translation and automatic evaluation methods (Banerjee and Lavie, 2005). Sentence-score is the relative evaluation between translations of the same source sentence. Shi et al. compared the rankings of machine translations of 300 source sentences (Shi et al., 2012). System-score is the evaluation score among machine translation systems in the given dataset of source sentences. Goto et al. scored 23 translation systems using 300 source sentences, and the average of these scores was treated as the score of the translation system (Goto et al., 2011). To verify the usefulness of crowdsourcing in evaluating translations by each score, we propose that the evaluation scores be compared. By using translation evaluations yielded by the following criteria, we compare professional evaluations to crowdsourced evaluations.

First, we define the evaluation scores to be compared. There exist m source sentences and n translation systems that translate each source sentence. $s_{ij}(crowd)$ indicates the evaluation score of the$j$-th ($1 \leq j \leq n$) translation system output of the$i$-th ($1 \leq i \leq m$) source sentence, which was made by crowdsource workers.

In this research, we used more than one worker for each evaluation. The score for each sentence is defined as the average of all the evaluation result. Also, $s_{ij}(professional)$ is the evaluation score of the bilingual professional. In the experimental dataset, only one professional has evaluated each translation. Details will be explained in Section 4.

In this analysis, the crowdsourcing score is calculated as the average of all workers ’ adequacy scores of each translation of each source sentence. $s_{ij}(professional)$ is the evaluation score output by the professional. Based on this definition, the three scores compared are given below.

Translation-score : Compares the translation-score using all source sentences and all translation outputs.

$$S_{translation}(e) = \{s_{11}(e), \ldots, s_{1n}(e), \ldots, s_{mn}(e)\} \quad (1)$$

$S_{translation}(e)$ displays the vector of all translations by evaluator e. Comparison of translation score is made in order to determine whether the absolute evaluation of the professional is the same as that obtained by crowdsourcing. The comparison of translation-score is based on the MAE (mean absolute error) value of each translation-score. A lower MAE indicates a closer score.

Sentence-score : Sentence-score uses the relative correlation among translation systems for one source sentence. Different from sentence-score, it is the correlation between crowdsourcing and professional for the translations of sentences. $S_{sentence}(i, e)$, which is the set of evaluations for source sentence i evaluated by e, is obtained as below:

$$S_{sentence}(i, e) = \{s_{i1}(e), s_{i2}(e), \ldots, s_{in}(e)\} \quad (2)$$

Sentence-score compares the ranking results of two evaluation methods for a given source. Correlation coefficient is adopted to compare sentence-scores between crowdsourcing and professional.

System-score : System-score comparison uses the arithmetic mean of each translation system for all source sentences. Namely,

$$S_{system}(e) = \{\frac{\sum_{i=1}^{m} s_{i1}(e)}{m}, \ldots, \frac{\sum_{i=1}^{m} s_{in}(e)}{m}\} \quad (3)$$

$S_{system}(e)$ contains the score for each translation system by averaging the score for all source sentences. System-score comparison can determine the usefulness of the ranking in machine translation systems. Along with sentence-score, the system-score comparison is based on correlation coefficients.

Based on the three scores we obtain the crowdsourcing score and evaluation. Evaluation metrics are based on adequacy and fluency, as detailed in Section 4.

4. Experiment Design

To analyze the feasibility of crowdsourcing evaluation, a Chinese-to-English translation experiment was conducted. An evaluation task of Chinese-to-English translation was created, and crowdsource workers performed the evaluation task. In this experiment, we used Amazon Mechanical Turk (AMT)[1] as the platform for crowdsourcing. AMT is one of the largest crowdsourcing platforms, and focuses mainly on microtasks. NTCIR-PATENT data set was used for this experiment. It was created for translating international patent sentences among three languages: Japanese, English, and Chinese (Goto et al., 2011). The data was taken from international patent description sentences, and they don ' t resemble each other too closely. Documents were translated by 23 translation systems and each translation was evaluated by a bilingual professional. The metrics for evaluation were adequacy and acceptability. Preliminary training evaluated 100 translations by three professionals, so that they have a common understanding toward evaluation. This experiment extracted 10 source sentence data and 23 translations for each source sentence. After extraction, we tasked a English native speaker with assigning fluency scores. Crowdsource workers performed the evaluation task for a total of 230 translations.

The reason of conducting Chinese to English translation is the number of workers. Number of Chinese-English bilingual worker is much larger than Japanese-English bilingual in AMT. Annual report of U.S. Immigration Enforcement Actions says the most Asian people migrates from China[2]. Preliminary experiment of Japanese-to-English translation evaluation task resulted in six workers in five days, which makes impossible to collect enough data. The purpose of this experiment is the translation evaluation by bilingual workers, so we offered the task that expected to hire the most number of workers.

First, crowdsource workers were instructed as below:

- Please read each Chinese sentence and its English translation.
- Please evaluate the translation in terms of adequacy (or fluency).
- Finally, please reply to the questionnaires.

Also, here are the criteria given to workers (White et al., 1994).

Adequacy: the degree to which information present in the original is also communicated in the translation.

Fluency: the degree to which the target is well formed according to the rules of Standard Written English.

Both adequacy and fluency have 5-range standards, which indicate better qualities in higher score. Table 1 provides an example of the adequacy evaluation task. The source sentence and machine translation are shown along with the scores of a crowdsource worker. Each task includes 23 translations. Each crowdsource worker performs 10 tasks for a total of 230 translations.

Workers see the Chinese source sentence and it ' s translations in English. The reference translation is not provided in the task. Therefore, workers are required to have a bilingual ability. The reason for conducting Chinese to English translation is the number of workers. The number of Chinese-English bilingual worker is much larger than Japanese-English bilingual in AMT. Annual report of U.S. Immigration Enforcement Actions says the most Asian people

[1] https://www.mturk.com/mturk/welcome

[2] https://www.dhs.gov/publication/immigration-enforcement-actions-2012

Table 1: Example of evaluation task

Source sentence	Machine translation	Adequacy	Fluency
障碍物的一个可行的实施方案为在鼓内的螺旋状障碍物(图1中未示出)。 Reference: One possible embodiment of an obstacle is a helically-shaped obstacle (not shown in Fig. 1) within the drum.	One possible embodiment is within the drum of an obstacle (not shown in Fig.1).	5 (All meaning)	5 (Flawless English)
		4 (Most meaning)	4 (Good English)
		3 (Much meaning)	3 (Non-native English)
		2 (Little meaning)	2 (Disfluent English)
		1 (None)	1 (Incomprehensible)

migrate from China . Preliminary experiment of Japanese-to-English translation evaluation task resulted in six workers in five days, which makes impossible to collect enough data. The purpose of this experiment is the translation evaluation by bilingual workers, so we offered the task that expected to hire the most number of workers.

In addition, the screening process was designed to detect and eliminate spammers in this experiment. Spammers are unfair workers or just program for earning money automatically. This research handled spammers ’ problem in two ways. One is the limitation of accessible area, and the other is the qualification test.

The former method limited the workers from the access from U.S. This is because the task requires English and Chinese knowledge, and there are few workers lives in China. Joel et, al explains 92% of AMT workers are from U.S. or India (Ross et al., 2010). Another reason is that preliminary experiment resulted in poor quality submission except by workers from U.S. From these reasons, we set up the access limitation.

The another method, the qualification test is held to filter crowdsource workers. If workers don’t get the enough score, they cannot perform the evaluation task. The qualification test is the same design as the real task assigned to workers. We extracted one source Chinese sentence and ten translations of it. Workers makes 5-range evaluations by adequacy or fluency. The qualification is based on the correlation between test result and professional evaluation. In this experiment, workers with the correlation 0.4 or higher are qualified. The reason for this score is the the requirement for the worker is not professional quality but the honesty and a certain degree of bilingual ability Also, this is because we wanted to employ more workers. The task requires both bilingual ability and patent-specified knowledge, but this task only checks the evaluation ability.

5. Analysis of Experiment

This section compares crowdsource workers and professional evaluation, and to analyze the experimental result. For this we assessed the quality of translations and translation systems by MAE or correlation by using the standards introduced in Section 3. For each standard of adequacy and fluency, we tried to identify the points of similarity and difference between crowdsourcing and professional evaluation.

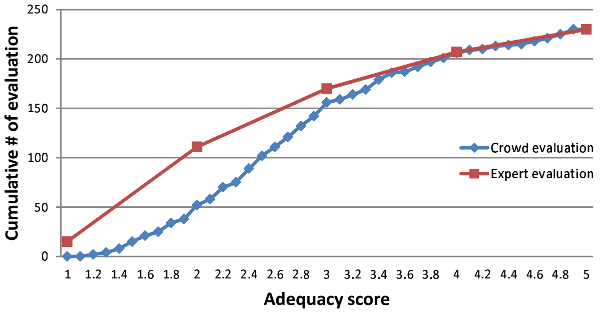

Figure 1: Cumulative graph of professional and crowd evaluation of adequacy

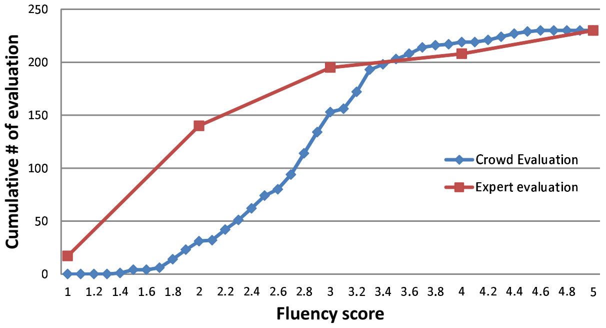

Figure 2: Cumulative graph of professional and crowd-sourcing evaluation of fluency

5.1. Translation-score Comparison

The first comparison is that of translation-score. This comparison addresses the absolute score. This comparison resulted in MAE scores of 0.63 and 0.68 for adequacy and fluency, respectively. Figures 1 and 2 plot the cumulative graph of professional and crowdsource worker evaluations of adequacy and fluency, respectively. These graphs show the large and small differences in professional scores and crowdsourcing evaluations. The average result of collected workers approached the professional assessment at the score of 3, the median of the 5 levels. Scores greater than 3 evidenced only a very small difference. This means that crowdsourcing workers tend to assign higher scores to low-quality translations. This tendency is most clearly obvious in the fluency scores. 56% of translations were scored between 2.5 and 3.5 (averaged). On the other hand, only 24% of translations were scored three by the professional. These results suggest that combining the results of multiple workers boosts the centralization of scores.

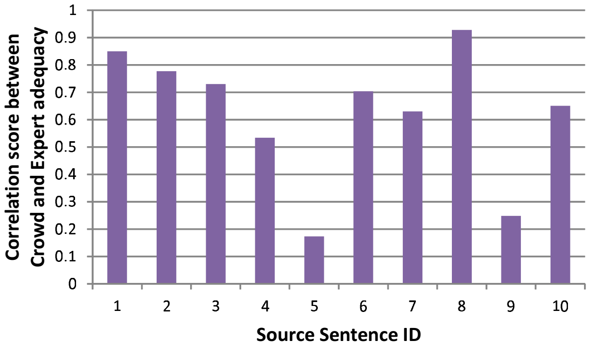

Figure 3: Bar graph of correlation for each source sentence by adequacy

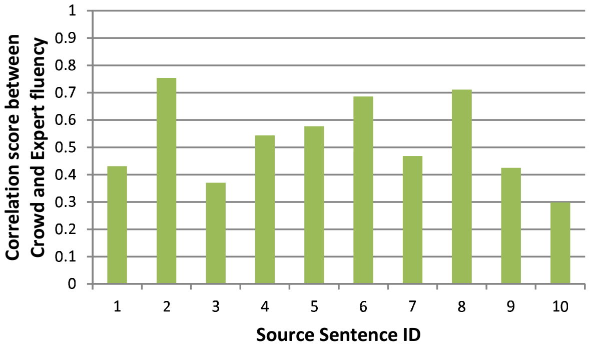

Figure 4: Bar graph of correlation for each source sentence by fluency

5.2. Sentence-score Comparison

The sentence-score comparison indicates the effectiveness of adequacy ranking in each source sentence. Figures 3 and 4 plot bar graphs of the sentence-score correlation; they show that the variation in evaluation score by the crowdsourcing workers is quite large for each source sentence. As for adequacy, the average of each sentence ’ s correlation is 0.62; maximum correlation of 0.93, and minimum correlation is 0.17. As for fluency, average of correlation is 0.53; maximum of 0.75 and minimum of 0.3. These wrong evaluations occur due to the difference in the evaluation criteria between crowdsourcing workers and professional. A detailed example of this is shown in Section 6. This problem occurred because of the closeness of professional evaluations. In source sentence 5, the professional scored 17 of 23 translations as 2. This unbalanced distribution in evaluation worsens the crowdsourcing evaluation result.

5.3. System-score Comparison

Finally, the system-score comparison showed the crowdsource workers ranked the 23 machine translation systems in agreement with the professional. This bar graph of system-score shows the crowdsourced and professional evaluations. The result was the correlation of 0.84 between crowdsource workers and the professional. This is the best result among the three comparisons. The best translation system as determined by the professional was also scored best by the crowdsource workers (See Figure 5). Moreover, according to Figure 6, fluency yields a similar result in selecting the best translation system. The correlation coefficient of system-score as regards fluency is 0.80 and the best translation system can be selected by crowdsourcing. Therefore, crowdsourcing is useful in selecting the best machine translation system. The reason for these good results may be the reduction in errors by the effect of averaging the number of translations.

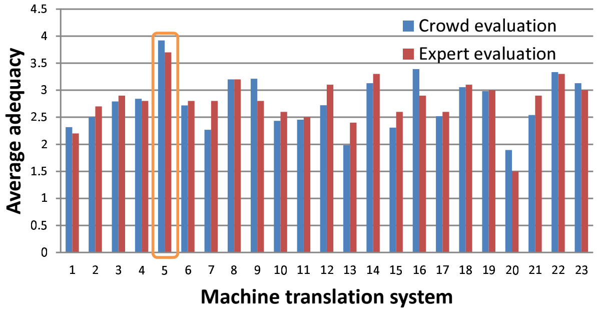

Figure 5: Bar graph for each translation system of adequacy. Crowdsourcing can select the best system, Machine translation system 5.

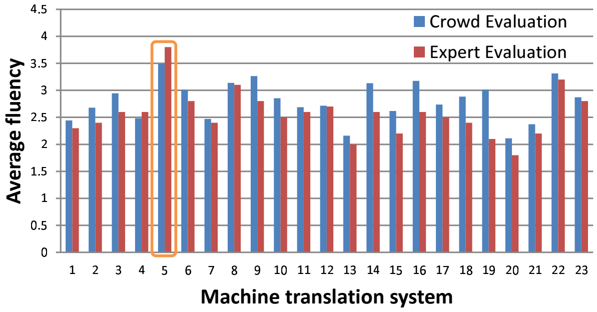

Figure 6: Bar graph for each translation system of fluency. Crowdsourcing can select the best system, Machine translation system 5.

Here ’ s a summary of our crowdsourcing analysis:

- From translation-score, crowdsourcing workers tend to assign higher scores than professional bilinguals. As for high-quality translations, crowdsourcing workers and professionals assign almost the same scores.
- From sentence-score, the correlation between crowdsourcing workers and professional varies widely with the source sentence.
- From system-score, crowdsourcing is useful as it offers high correlation and can identify the best translation system.

Translation-score and system-score exhibited little difference between adequacy and fluency. The difference between adequacy and fluency according to crowdsourcing

Table 2: Example of different evaluation professional and crowdsourcing by adequacy

Source sentence	Machine translation	Professional adequacy	Crowd adequacy
可以理解，上述系统300确保无论何时客户端应用程序实例化模式化类型的实例以便保存在存储中，该客户端就能访问模式包。	It is to be understood that the above-described system 300 ensures that whenever a client application instanciation modes, examples of the types to be stored in the memory, the client can access mode.	2	3.5
Reference: It is to be understood that the above described system 300 ensures that whenever the client application instantiates an instance of a schematized type for persistence in a store, the client has access to the schema package.	It is to be understood that the above-described system 300 to ensure that whenever the client application **instantiated schematized** types of **Examples** to stored in storage , the client can access mode packets .	5	2.83

Table 3: Example sentences of different evaluation between professional and crowdsourcing by fluency

Source sentence	Machine translation	Professional fluency	Crowd fluency
因为X3和X4不穿过信号变换单元，所以各自的声道配置信息为0。	Since X3 and X4 does not pass through the signal conversion unit, so that respective channel configuration information 0.	1	3.22
Reference: Since X3 and X4 do not pass through signal converting units, each channel configuration information becomes 0.	Because X3 and X4 do not pass through the signal translation unit, therefore the respective sound track configuration information is 0.	4	2.6

was biggest in the sentence-score evaluation. This is due to the impact of source sentence; the professional had a wider score distribution. In Section 6, we explain how this difference occurred.

6. Discussion

The result of the comparison between crowdsourcing and professional evaluation showed the feasibility of system score, and showed the difficulties of utilizing sentence-score and translation-score. This section explains a specific example of useful and useless translation to validate the use case of crowdsourcing, In addition, we refer to the cost and time for crowdsourcing for practical use.

Comparison between Adequacy and Fluency

We display some examples to show the difference between crowdsourcing evaluation and professional evaluation in Table 2. The machine translation omits some information of the reference translation such as "schematized" and "package". However, crowdsource workers gave higher scores then the professional because the translation is grammatically correct. On the other hand, the machine translation result below was scored 5 by the professional, and the average of the crowdsource workers was 2.83. The workers focused on the grammatical errors even though they had little impact on adequacy (Bold type in the sentence).

As for fluency, the upper translation of Table 3 was evaluated as "Incomprehensible" by the professional. This is because the later translation clause lacks a verb. However, crowdsourcing workers took the prior correct expression into account, so the average fluency was 3.22. The lower translation was scored 4 by the professional. However, the crowdsourcing workers considered the mistranslation of " sound track " , so the fluency score was 2.6.

From these examples, the instructions given to the workers should distinguish adequacy and fluency explicitly, to eliminate their misunderstanding. Also, the workers were negatively impacted by their lack of knowledge about the patent domain. Future improvement could be achieved by combining monolingual workers who have patent knowledge and bilingual workers who have language knowledge. Also, towards the practical use of translation evaluation by crowdsourcing, a discussion about cost and time is needed in addition to quality.

Cost and Time

In this experiment, two dollars were spent to evaluate 23 translations. This equates to an hourly rate of 18 dollars per worker. A survey reported that a questionnaire task that required no special ability cost about 1.71 dollars (Paolacci et al., 2010). Compared to this task, our experiment is more costly. This is because we employed more workers regardless of cost. For a comparison with professionals, the Japan Association of Translators 3 says that English to Japanese translation [3] costs 3,000 to 10,000 yen (about 30 to 100 dollars) per page.

As for consumed time, we collected 14 workers during the two week experiment. This is mostly because of the difficulty of the qualification test. Also, time cost is affected by various factors. Examples are monetary cost, translation language pairs, and the difficulties of the qualification test. As mentioned in Section 4., most workers accessed from the U.S.A. or India, so few workers will participate in tasks

[3] http://jat.org/working_with_translators/

in which both source language and target language are non-English. To address this problem, other approaches such as a different crowdsourcing platform are needed.

Also, the workers ’ participation time had some effect on the result. Average evaluation time was 7 minutes for the adequacy evaluation and 4.5 minutes for the fluency evaluation. This difference occurred because of one worker who took an inordinately long time. That worker took 20 minutes on average. We found that shorter completion times were matched by higher correlation in both adequacy and fluency.

Guidelines on crowdsourcing language resources

Here are some initial guidelines for utilizing crowdsourcing to create a translation evaluation corpus. First, we should prepare “ good ” source sentences, because poorly-written source sentences confuse most workers. Of particular note, this task requires bilingual ability and there are no reference translations by professional bilinguals, so the quality of the source sentence will have a strong impact on task quality.

In addition, as for worker ability, the minimum requirement is to be a native speaker of either language. We need qualification tests for bilingual ability that are easier to complete and confirm.

If crowdsourcing evaluation is applied to the feedback of translation, the task design is a considerable solution. An example can be binary check of the translation quality with “good” or “bad”, and the relative ranking for each translation.

7. Conclusion

Various methods have been proposed to evaluate translation quality. However, existing methods have problems in terms of cost and/or quality. In this study, we have adopted the crowdsourcing approach to achieve high-quality and low cost evaluations. We conducted three comparisons to determine the practical area of crowdsourcing evaluations. We showed three comparison schemes for translation evaluation based on the usage of translations: translation-score, sentence-score, and system-score. Comparisons by these scores were used to assess the absolute evaluation of translated sentences, relative evaluation of source sentences, and relative evaluation of machine translation systems. In addition, comparison of crowdsourcing scores and professional scores found that crowdsourcing evaluations are useful when the goal is to determine the best machine translation system given multiple source sentences and translation outputs.

In addition, based on comparison schemes, we conducted the crowdsourcing evaluation experiment. A comparison of crowdsourcing scores and professional scores found that crowdsourcing evaluations are useful when the goal is to determine the best machine translation system given multiple source sentences and translation outputs.

Future work includes developing methods to enhance the translation-score comparison or sentence-score comparison. For this, we will focus on the effect of the number of workers and the behavior of the crowd. In addition, an analysis of the effect of reducing the evaluation size of the task remains to be done. We have to make the crowdsourcing performance more efficient and more effective.

8. Acknowledgements

This research was partially supported by a Grant-in-Aid for Scientific Research(S) (24220002, 2012~ 2016) from Japan Society for the Promotion of Science (JSPS), and Service Science, Solutions and Foundation Integrated Research Program from JST RISTEX.

9. References

Ambati, V., Vogel, S., and Carbonell, J. G. (2010). Active learning and crowd-sourcing for machine translation. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC’10)*, volume 11, pages 2169–2174.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Bentivogli, L., Federico, M., Moretti, G., and Paul, M. (2011). Getting expert quality from the crowd for machine translation evaluation. In *Proceedings of the MT Summmit*, volume 13, pages 521–528.

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. (2010). Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 313–322. ACM.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, pages 249–256.

Callison-Burch, C. (2009). Fast, cheap, and creative: evaluating translation quality using amazon’s mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, volume 1(1), pages 286–295. Association for Computational Linguistics.

Chen, K. T., Chang, C. J., Wu, C. C., Chang, Y. C., and Lei, C. L. (2010). Quadrant of euphoria: a crowdsourcing platform for qoe assessment. *Network, IEEE*, 24(2):28–35.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Goto, I., Lu, B., Chow, K. P., Sumita, E., and Tsou, B. K. (2011). Overview of the patent machine translation task at the ntcir-9 workshop. In *Proceedings. of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologiess*, volume 9, pages 559–578.

Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 439–448. ACM.

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. ACM.

Lin, D., Murakami, Y., Ishida, T., Murakami, Y., and Tanaka, M. (2010). Composing human and machine translation services: Language grid for improving localization processes. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, volume 10, pages 500–506.

Little, G., Chilton, L. B., Goldman, M., and Miller, R. C. (2010). Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 57–66. ACM.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Ross, J., Irani, L., Silberman, M., Zaldivar, A., and Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 2863–2872. ACM.

Shi, C., Lin, D., Shimada, M., and Ishida, T. (2012). Two phase evaluation for selecting machine translation services. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1771–1778.

White, J., O'Connell, T., and O'Mara, F. (1994). The arpa mt evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas*, pages 193–205.

Zaidan, O. and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229.

Zhang, Y., Vogel, S., and Waibel, A. (2004). Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of the Forth International Conference on Language Resources and Evaluation (LREC'04)*, pages 2051–2054.