Master Thesis

Using Crowdsourcing for Evaluation of Translation Quality

Supervisor Professor Toru ISHIDA

Department of Social Infomatics
Graduate School of Informatics
Kyoto University

Shinsuke GOTO

February 7, 2013

Using Crowdsourcing for Evaluation of Translation Quality

Shinsuke GOTO

Abstract

In recent years, a wide variety of machine translation services have emerged due to the increase on demand for multilingual communication supporting tools. Machine translation services have an advantage in being low cost, but also have an disadvantage in low translation quality. Therefore, there is a need to evaluate translations in order to predict the quality of machine translation services. In most cases, quality of machine translation services is calculated by professionals using a 5-scale evaluation. However, evaluation by bilingual professionals requires a lot of time and money, which makes this evaluation approach unpractical to be used for all translation services.

In this study, we introduce a crowdsourcing translation evaluation method. Crowdsourcing is a novel framework where small tasks are completed by an anonymous crowd by request. The framework has advantages in task performance in terms of time and cost compared to professionals, but the language abilities of crowdsourcing workers cannot be guaranteed. Moreover, users aiming to gain illicit rewards from a task, such as spammers, exist among the crowdsourcing workers. Due to these problems, it is unknown whether the same quality can be obtained from crowdsourcing as from professional evaluators. The following problems need to be solved in order to utilize crowdsourcing evaluation of translation.

Feasibility of crowdsourcing evaluation of translation

It remains unclear whether crowdsourcing evaluation can replace professional evaluation. The error range between evaluation scores from crowdsourcing workers and professional contributors needs to be analyzed to determine feasibility of the crowdsourcing approach.

Change in evaluation quality by number of crowdsourcing workers

In general, a translation evaluation score approaches a constant score as the number of workers increases. There is a need to predict the convergence of evaluation score by crowdsourcing evaluation of translation.

In this research, an experiment using the Amazon Mechanical Turk (MTurk), the largest platform of crowdsourcing, was conducted. A Chinese to English translation evaluation task of 5-range evaluation was designed, workers were asked to perform the task for a 0.05 dollar compensation for each evaluation. In order to filter out malicious workers, a qualification test was introduced.

Secondly, the crowdsourcing evaluation of translation was analyzed regarding the following aspects:

Comparison of individual evaluation by crowdsourcing workers and professional contributors for each given translation. The individual evaluation by crowdsourcing workers resulted in the large disparity compared to the evaluations from professionals. At the same time, the effect of increasing the number of workers to number of errors compared to professionals was evaluated. The scores were calculated by using various aggregation methods: average, median, and mode. The result suggests that median of all evaluation results can be a low-cost and high-quality evaluation.

Finally, an application to for estimating translation quality by crowdsourcing evaluation was implemented. This system consists of a task request module and an evaluation management module. The evaluation management module shows the estimated quality of crowdsourcing evaluation, which enables requesters to determine whether the evaluation result is feasible or not.

Contributions of this research are as follows:

Feasibility of crowdsourcing evaluation of translation

The median score of crowdsourcing evaluation has a 0.28 mean absolute error compared to professional evaluation. Therefore, crowdsourcing evaluation quality is high enough in the translation evaluation area.

Change in evaluation quality by number of crowdsourcing workers

In this research, 30 workers evaluated the translation quality for each given sentence. Then, the effect of increasing the number of workers on the changes of evaluation score was analyzed. The results showed the median of crowdsourcing evaluation can minimize the difference with professional evaluation. As a result, the least difference to professional evaluation at 0.28 was found to be the median for 30 crowdsourcing evaluations.

クラウドソーシングを用いた翻訳品質の評価

後藤 真介

内容梗概

近年，多言語コミュニケーションの重要性の高まりに合わせ，多くの機械翻訳サービスが生み出されている．機械翻訳サービスによる翻訳は人手による翻訳に比べてコストが低い点が特徴であるが，翻訳精度が保証されない問題がある．そこで，翻訳品質を予測するために評価が必要となる．多くの場合，機械翻訳サービスの評価のためには多くの文章に対して専門家が 5 段階評価を行い，その平均値によって翻訳精度の予測値が取得される．しかしながら，専門家による評価は多くの時間的コストと金銭的コストを必要とし，全ての機械翻訳サービスに対して評価を行うことは容易ではない．

そこで，本研究では翻訳を評価する手法としてクラウドソーシングを導入する．クラウドソーシングはタスクを不特定多数の作業者に依頼するシステムであり，タスクを安価かつ迅速に実行可能であるという利点がある．しかしながら，クラウドソーシングの作業者は語学力が保証されない．さらに，スパマーと呼ばれるタスクに対する報酬を不正に受け取ろうとする作業者も存在するため，専門家と同等の品質の評価がクラウドソーシングによって得られるかどうかは分からない．また，クラウドソーシングは低コストといえど無料ではなく，特に多くの文章や機械翻訳に対する評価が必要となる場合は適切な作業人数を決定する必要がある．

このため，翻訳の評価をクラウドソーシングによって行う際には以下の点が課題となっている．

クラウドソーシングによる翻訳評価の有効性分析

クラウドソーシングによる翻訳評価には上述の課題があり，専門家の評価の代替として使えるかどうかは明らかになっていない．クラウドソーシングの作業者による評価結果が専門家の評価値に近い評価であるのかを分析する必要がある．

クラウドソーシングの作業者数による翻訳評価値の品質の変化

一般的に，評価データ数の増加に従って集計後の評価値は一定の値に収束する．今回，翻訳品質の評価にクラウドソーシングを適用するにあたり，作

業人数の増加による翻訳評価値の収束を分析する必要がある．

本研究では，世界最大のクラウドソーシングプラットフォームである Amazon Mechanical Turk(MTurk) を用いた翻訳評価実験を実施した．実験においては，中英翻訳の評価を作業者に依頼した．評価基準として専門家と同じ 5 段階評価を作業者に指示し，報酬として各翻訳文の評価に対し 0.1 ドルの報酬を設定した．また，専門家の評価値との類似度を測定する式として平均絶対誤差を用いた．

クラウドソーシングによる翻訳評価に対し，以下の方針で分析を行った．まず，作業者数の増加によって専門家の評価との誤差がどのように変化するかを調べた．同時に，複数人の作業結果の集計方法として平均値，中央値，最頻値のそれぞれを用いた場合に最も専門家の評価に近い集計方法がどれであるかを検証した．分析の結果，作業者数の増加によりの専門家の評価との平均絶対誤差は減少することが確かめられた．また，集計方法について，専門家の評価値によって専門家の評価と類似した集計法は異なることが判明した．専門家が高く評価を下した翻訳文に対しては作業者の評価結果の中央値が専門家に近い評価値であり，専門家の評価値が 5 段階評価の中心である 3 点の場合は，作業者の評価結果の平均値が専門家と類似した評価を行う分析結果が得られた．

最後に，クラウドソーシングを用いた翻訳評価の実用化に当たり，クラウドソーシング翻訳評価システムの設計を行った．本システムは評価依頼モジュールと評価結果管理モジュールからなり，評価結果管理モジュールではクラウドソーシングによる作業結果の評価品質を予測することにより依頼者が翻訳評価結果を利用可能かどうか容易に判断可能である．

本研究の貢献は以下の通りである．

クラウドソーシングによる翻訳評価の有効性分析

クラウドソーシングによって翻訳評価を行った場合，全員の評価値の中央値を真の評価値の予測として扱った結果，専門家の評価値との平均絶対誤差は 0.28 であった．この結果から，クラウドソーシングによる翻訳評価は専門家と比べても十分な評価品質を持っていることが分かる．

クラウドソーシングの作業者数による翻訳評価値の品質の変化

今回，各翻訳文に対して 30 人の作業者が翻訳品質評価を行った．そこで，それぞれの結果に対し，作業人数が増加することによる翻訳評価値の変化の分析を行った．その結果，人数が増えるほど評価値の分散は小さくなり，また専門家の評価値との誤差が小さくなる知見を得た．

Using Crowdsourcing for Evaluation of Translation Quality

Contents

Chapter 1 Introduction

Recently, more and more machine translation services are becoming available. For example, Language Grid, an infrastructure for services computing, offers various language services like multilingual dictionaries and machine translations with the interface of Web services [1]. In addition, language services can improve the quality of translation by combining bilingual dictionaries. In the case of translating the document for sightseeing in Kyoto, machine translation services can be combined with the dictionary of sights in Kyoto. Due to these combined translation services there are increasing demands for evaluation of translation services with the low cost and high quality. If the low-cost evaluation is available, best service in each situation can be selected dynamically so that users of language services can enjoy these services. Also, NTCIR-9 PATENT held the evaluation task for patent document [2]. 18 machine translation systems are submitted translation data for the test sentences for measuring the translation quality of machine translations.

Traditionally, quality evaluation of machine translation has been held by bilingual professionals. In most situations, translation quality is assessed by adequacy, which is the criterion that refers to the degree to which the translation communicates information present in the original text. Evaluating adequacy of many translations by 5-range, the average score resulted to be the adequacy of the translation service. However, these evaluations by professionals are costly, and this makes it not feasible to evaluate all the translations for each translation services. For example, Koehn determined the adequacy of machine translation by gathering the evaluation of 200 to 300 sentences from more than 30 people [3]. In addition, machine translation evaluation in NTCIR-9 evaluates 300 task sentences for each machine translation system. In this task, three professionals got training by evaluation of 100 translations before actual evaluation.

On the other hand, kinds of methods are focused on automatic evaluation of translation. For example, BLEU and METEOR are well-known automatic evaluation methods based on the similarity between machine translation and reference translation [4][5]. These methods have good correlation with pro-

fessional, but are difficult to obtain absolute quality of translation. Another problem is that these automatic evaluation methods requires one or more reference translations for calculating similarity. If the domain changes dynamically, even automatic evaluation requires high cost to get reference translations. These facts necessitate the low-cost, fast, and high-quality evaluation method for translation.

In this study, we introduce crowdsourcing evaluation as the low-cost evaluation that have same quality as professional evaluation. Crowdsourcing is the new marketplace for employers and emproyee in which they can request or perform task. In addition, crowdsourcing have the characteristics in the low cost and open to the Web. Comparing to the professional work, crowdsourcing can be quick and low-cost. This is one of the new form of collective intelligence, and is expected to achieve significant growth. Also, crowdsourcing can collect workers from wide range of countries due to the unrequired face-to-face contact. This feature has large merit in translation evaluation task according to the translation locality. .

Our experiment was conducted in Amazon Mechanical Turk (MTurk) [1)] for crowdsourcing platform. MTurk is the largest crowdsourcing platform in the world. Each task is called HITs (Human Intelligence Tasks) in MTurk, and currently 2,900 kinds of HITs group and 270,000 HITs are registered by task requester as of February of 2013. These HITs are performed by crowdsourcing task workers (hereinafter referred to as worker) from over 100 countries. For example, dictation from speech data and annotation of graphical data are allocated. These tasks are hard to be done by machines, but they are very low cost in MTurk.

However, the quality of workers in crowdsourcing is generally not guaranteed. Major tasks in crowdsourcing don't require workers of special ability, and quality of these results is compensated by redundancy in most cases. Redundancy is achieved by rules of majority, so more than a half of workers should perform correctly. Also, there are spammers, who are aiming to gain illicit re-

[1)] http://www.mturk.com

ward, in the crowdsourcing market. Ipeirotis *et al.* noted 30% of workers are spammers, which prevent requesters from obtaining high-quality results [6]. For those reasons, evaluation of translation, which requires translation ability to a certain degree, is not held in crowdsourcing.

This research focused to analyze the feasibility of crowdsourcing evaluation of translation. If crowdsourcing evaluation turned out to be viable, it can be replaced with the professional evaluation. We deal with the issues below:

Feasibility of crowdsourcing evaluation of translation

As mentioned above, crowdsourcing evaluation is not clear to have feasibility to replace professional evaluation. We should alanyze for the assessment of the similarity and difference of crowdsourcing and professional evaluations. Also, we have to clarify the problems in crowdsourcing evaluation of translation, and the feasible evaluation strategy for crowdsourcing.

Change of evaluation quality by number of crowdsourcing workers

In general, the evaluation score gets convergent as the number of evaluation increases. However, specific number of evaluation for the appropriate score is not discussed. This research need to analyze the form how to convergent the evaluation score as the increasing workers.

To address these issues, we conducted the experiment using crowdsourcing, we will show the feasibility of crowdsourcing evaluation in real crowdsourcing marketplace.

Firstly, we measured the evaluation quality of workers, and then analyzed the difference and similarity between crowdsourcing evaluation and professional evaluation. Secondly, we validate the effect of the aggregation method in increase number of evaluation, by using errors from professional score as evaluation metrics. Finally, utilizing knowledge from analysis, we propose the interface that makes full use of MTurk by time and cost, and introduce the prototype of the system.

Figure 1 is the concept diagram of this research. This shows the translation evaluation is determined by three criteria: cost, speed, and quality. Also, there are trade-off among three criteria; in one hand, professional evaluation has very high quality at the expense of cost and speed. On the other hand, automatic

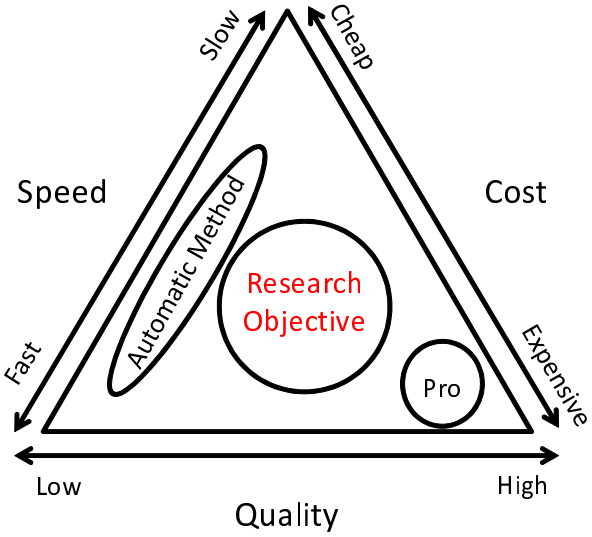


Figure 1: Trade-off of each criteria for translation evaluation

evaluation has low quality in return for its good cost and speed. In this research, we analyze the feasibility of new method of crowdsourcing evaluation, and estimate whether it is useful for real field.

This paper is organized as follows: In Chapter 2, we introduce the crowdsourcing research and translation evaluation method. These researches have both the relation and differences than our research. Next, in Chapter 3 clarify the purpose in crowdsourcing we will use, and we explain the experimental method we conducted in MTurk. We analyze the result of experiment in Chapter 4, comparing the result of crowdsourcing evaluation and professional evaluations, and then propose the interface using obtained knowledge in Chapter 5. Finally, we discuss the result of crowdsourcing in Chapter 6, and Chapter 7 concludes this paper.

Chapter 2 Related Work

There are three main areas as our related works: researches related to the evaluation method for translation, those about crowdsourcing, and those of crowdsourcing evaluation of a kind of metrics.

2.1 Translation Evaluation Method

Until recently, evaluations of translation quality are collected by man-hand. In many cases, translation quality is estimated by adequacy and fluency [7]. Adequacy refers to the degree how proper the translated sentences conveys correct meaning, and fluency refers to the how correct the translated sentence is as the translated language. Also, some evaluation schemes are based on order relation [8]. In this scheme, professionals rank multiple translations to obtain the best machine translations. These evaluation methods by man-hand requires professional and long time to complete.

In addition, various kinds of automatic evaluation methods have been proposed. Remarkable examples are BLEU and METEOR [4][5]. The basic idea of these evaluation methods is the similarity calculation by n-gram. These researches show the high correlation between professional evaluation and automatic evaluation, but similarity calculation based on n-gram is only applicable for same language. In most cases in automatic evaluation requires reference translation to compare with the target translation.

Gamon *et al.* proposed another automatic evaluation method without reference translation [9]. Their method is to calculate fluency by SVM and language model. However, this method is less better then n-gram based methods. This method evaluated only translated sentence, so adequacy is not estimated. These past researches show the evaluation of translation is very difficult in the cases without reference translations.

This research focused on low-cost and fast evaluation with professional quality. We analyze crowdsourcing evaluation without reference translation for achieving this objective.

2.2 Crowdsourcing

Crowdsourcing, the new working and employ style, requests tasks to anonymous workers, which can dramatically reduce the human cost. There are many kinds of crowdsourcing. One is the just web-based outsourcing, such as designing Web site or other complex works. InnoCentive [1)] and Crowdspring [2)] are the remarkable examples of this type of crowdsourcing. It is based on worker matching, the difference from existing work is the way to collect workers. The contents of work is the same, the reward is based on contracts.

Another type is that called microtasks, which this research focused on. Crowdsourcing by microtask requires task with very low cost, less than one dollar in most cases, and the large number of tasks are registrered. For example, dictation from pictures and relevance evaluation of Web search engine are conducted in microtask marketplace. The requesters focused on performing a large number of simple tasks, these tasks are previously conducted by a volunteer workers. Microtask outcomes the problem of time and place, and then it becomes the successful platform in commerce.

Especially, we give a deep explanation about Amazon Mechanical Turk (MTurk) as the microtask platform, which we use for crowdsourcing experiment. MTurk has two kinds of participator: requesters and workers. Requesters register to MTurk the tasks that are difficult for computers but easy to be solved by humans. In MTurk, tasks are called Human Intelligent Tasks (HITs). Workers performs HITs the requesters registered, and gain rewards from the result. Requesters can give the bonus to good workers, and instead, refuse the reward from result or block the worker. In MTurk, typical HITs are annotation of pictures or dictation of speech. In most cases HITs doesn't requires special ability.

Figure 2 is the screenshot of MTurk task selection. MTurk provides lots of available HITs that can be performed in ten minutes or around. Workers cant select any task and they will perform from those tasks. Figure 2 includes HITs such as the error check of transcript or baseball video annotation qualification.

[1)] https://www.innocentive.com

[2)] http://www.crowdspring.com

Figure 2: Amazon Mechanical Turk : Task Selection Screen

Also, each HIT description shows the Qualification for workers and the reward for each HIT. In most cases, rewards for each task are less than 0.1 dollar; they have to perform a lot of tasks to earn sufficient income.

The demographics of MTurk workers is described by Gabriele *et al.* and Ipeirotis in detail [10][11]. The results show that the largest percentage of workers are from US, which is about 47% . Currently, workers from India has been increased by 34%. Also, Gabriele shows the demographics of gender and education level in US.

In the US, female workers (64%) are more in number than male workers, and average age of workers are 36 years old, which is a little younger than average age of US population. Education level of MTurk workers is higher than average in US. More than 50 % workers are university graduated. This is caused by the demographics of start-up Web service user: recently, another survey shows worker of the lower education and life level is increasing.

Moreover, most of workers (69%) performs HITs as just killing time and fun.

Some workers, on contrary, performs more than thousand HITs to earn 1000 dollars. This statistics shows the MTurk worker in US have high education level, which we expect the workers with proper ability for translation evaluation task. The object of this experiment has the surveying whether the workers available for evaluation tasks are in crowdsourcing marketplace.

Ipeirotis also summarize the complete time in crowdsourcing tasks [12]. He pointed out that the complete time, the interval from task request to complete, is determined by exponential distribution, so the predict of accurate time is difficult. From the distribution, there are two bursts of completing HITs: 10 hours and around 170 hours. This is because the cases when requesters determine the task complete time as round time that number of HIT group increase around 10 hours. Also, increase number around 170 hours shows the standard task duration is set as one week (= 168 hours). About 90% of tasks are completed less than one week, whereas some tasks are unfinished until 1,000 hours. Based on these results, this research assessed complete time of the translation evaluation task.

2.3 Crowdsourcing Evaluation of Various Metrics

Crowdsourcing is used for various kinds of quality evaluation for standards. A good example is the research of Chen *et al.* that measures the QoE (Quality of Experience) of network by crowdsourcing [13]. Workers evaluate the subjective quality of multimedia contents by pairwise comparison. The analysis is conducted by comparing the crowdsourcing with the volunteer evaluators.

Voting by crowdsourcing is also regarded as evaluation. For example, voting is sometimes performed during the complex workflow by the crowdsourcing. Turkit prepares the voting function to carry out iterative improvement of quality [14]. Turkit enables the annotation of photo by long-text and the evaluation of multiple texts.

As explained above, crowdsourcing evaluation is executed by a wide range of fields, but up to now, not many papers allocated workers to the task with profession. The recent work, which is a workshop paper, Luisa *et al.* evaluated crowdsourcing evaluation of translation [15]. They asked workers to perform

the task to evaluate translations relatively: shows one source sentence and five translations and then make an order to them. Also, as the quality management, they used the gold-unit test, which is the question with obvious answer. If a worker submit a wrong answer to the gold-unit test, he will be rejected. Using CrowdFlower [1)] as the crowdsourcing platform, 52 workers are collected in 6 days. As a result, the proposed method has much better Spearman's correlation score than automatic evaluation method like BLEU and NIST. The objective of this research is not obtaining the order relation among machine translations, but the absolute 5-range evaluation quality.

Our work focused on comparison between professional bilingual and crowdsourcing worker. We conduct an 5-range evaluation experiment and compare crowdsourcing evaluation with professional by absolute error.

[1)] http://crowdflower.com

Chapter 3 Translation Evaluation Experiment by Crowdsourcing

This research conducted the translation evaluation experiment using crowdsourcing, in order to measure the feasibility of crowdsourcing. The experiment is aimed at measure the evaluation quality of workers, and analysis of improvement of evaluation quality by the number of workers. So designing proper task of crowdsourcing and managing quality are very important to improve performance.

In this chapter, we describe the task and quality control in the experiment in MTurk. Also, we mention the method to control the quality of workers.

3.1 Evaluation Criteria of Translation

Table 1: Adequacy Criteria

Score	Adequacy
5	All
4	Most
3	Much
2	Little
1	None

Table 2: Fluency Criteria

Score	Fluency
5	Flawless
4	Good
3	Non-Native
2	Disfluent
1	Incomprehensible

This time, the workers are required to make a 5-range evaluation of translations by adequacy and fluency. Here, we will explain the result of evaluation. We used adequacy and fluency as evaluation standards [16].

Table 1 and Table 2 shows the evaluation standard of adequacy and fluency. Adequacy indicate the degree to which information present in the original sentence is also communicated in the translation. Fluency refers to how well-formed translated sentences are according to the translated language. In the context of machine translation, adequacy is put more priority than fluency. In Chapter 4 we compare adequacy of crowdsourcing with professionals. This is because fluency is determined by only translated sentence.

Based on these definitions, workers are asked to read the instruction about adequacy and fluency, and make 5-range evaluations to each translations.

In this work we calculate 5-range evaluation as the interval scale. This is because most of related works about evaluation of translation and other areas use average and variance to the evaluation.

3.2 Experiment Design

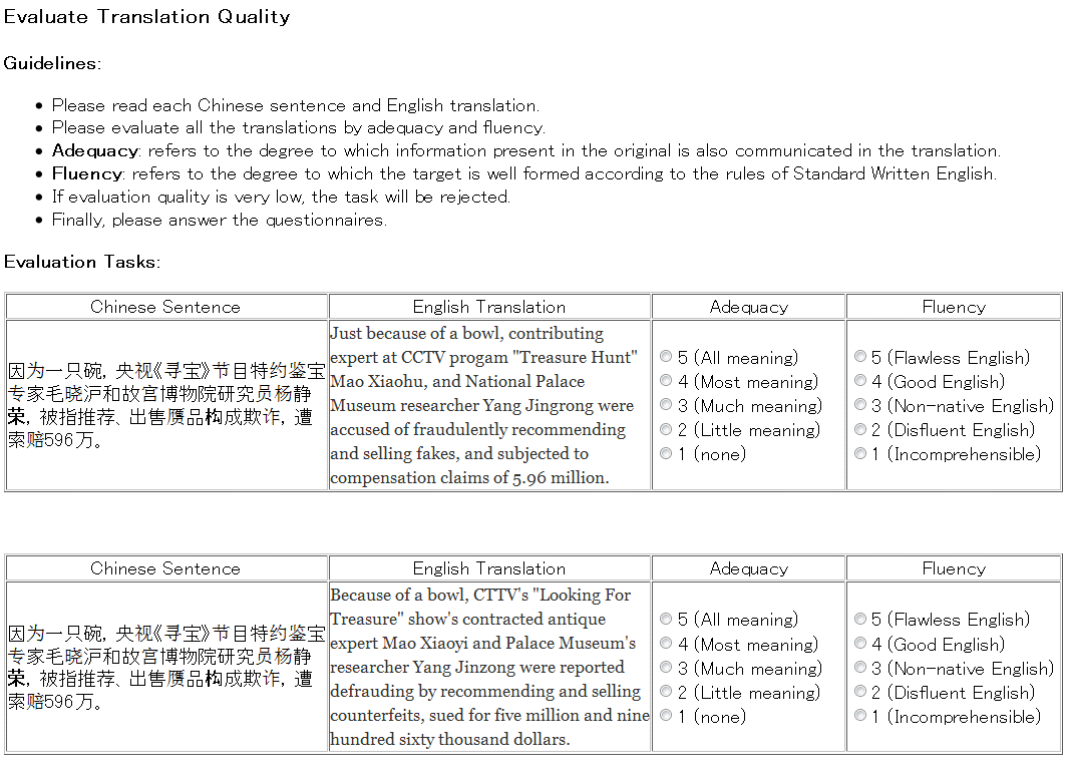
In this section, we explain the task the workers performed in detail. This experiment is Chinese to English translation we analyze the feasibility of crowdsourcing evaluation. Chinese source sentence is extracted from articles in Jingbaowang (京报网) [1], one of the most famous Chinese news site. English translation sentences are created by other tasks in Amazon Mechanical Turk. Fourteen translations are collected by Evaluation scores by professionals are obtained by three Chinese-English bilingual foreign students from China. Professional evaluation is the median of three.

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 MTurk. Preliminary experiment result in six workers in five days, which makes impossible to collect enough data. Details are in the Chapter 4, but the collection of workers are much better in Chinese-English translation evaluation.

Figure 3 shows the screenshot of 5-range evaluation. This experiment use MTurk, and the conducted the same standards as professionals. The instructions to the workers are below:

- Please read each Chinese sentence and English translation.
- Please evaluate all the translations by adequacy and fluency.
- Adequacy: refers to the degree to which information present in the original is also communicated in the translation.
- Fluency: refers to the degree to which the target is well formed according to the rules of Standard Written English.

[1] http://www.bjd.com.cn/



Evaluate Translation Quality

Guidelines:

- Please read each Chinese sentence and English translation.
- Please evaluate all the translations by adequacy and fluency.
- **Adequacy**: refers to the degree to which information present in the original is also communicated in the translation.
- **Fluency**: refers to the degree to which the target is well formed according to the rules of Standard Written English.
- If evaluation quality is very low, the task will be rejected.
- Finally, please answer the questionnaires.

Evaluation Tasks:

Chinese Sentence	English Translation	Adequacy	Fluency
因为一只碗，央视《寻宝》节目特约鉴宝专家毛晓沪和故宫博物院研究员杨静荣，被指推荐、出售赝品构成欺诈，遭索赔596万。	Just because of a bowl, contributing expert at CCTV progam "Treasure Hunt" Mao Xiaohu, and National Palace Museum researcher Yang Jingrong were accused of fraudulently recommending and selling fakes, and subjected to compensation claims of 5.96 million.	5 (All meaning) 4 (Most meaning) 3 (Much meaning) 2 (Little meaning) 1 (none)	5 (Flawless English) 4 (Good English) 3 (Non-native English) 2 (Disfluent English) 1 (Incomprehensible)

Chinese Sentence	English Translation	Adequacy	Fluency
因为一只碗，央视《寻宝》节目特约鉴宝专家毛晓沪和故宫博物院研究员杨静荣，被指推荐、出售赝品构成欺诈，遭索赔596万。	Because of a bowl, CTTV's "Looking For Treasure" show's contracted antique expert Mao Xiaoyi and Palace Museum's researcher Yang Jinzong were reported defrauding by recommending and selling counterfeits, sued for five million and nine hundred sixty thousand dollars.	5 (All meaning) 4 (Most meaning) 3 (Much meaning) 2 (Little meaning) 1 (none)	5 (Flawless English) 4 (Good English) 3 (Non-native English) 2 (Disfluent English) 1 (Incomprehensible)

Figure 3: 5-Range Evaluation Task in MTurk

- If evaluation quality is very low, the task will be rejected.
- Finally, please answer the questionnaires.

The most important point in the instruction is to give the definition of adequacy and fluency to the workers. The definition of adequacy and fluency is general for professional, they can evaluate without the explanation in detail. However, this may be the first evaluation of translation by a crowdsourcing worker, and in most cases they are not used to these evaluation method. So we firstly give the definition to make evaluation tasks easily to the crowdsourcing workers.

As for reject of result, this experiment doesn't reject any submitted data. The one of instruction just says the possibility to the reject. This field of research, Ipeirotis *et al.* explain in detail [6]. Massive evaluation experiment, this may be important in the future.

For each task, ten translations are allocated, and the reward paid for each task is 0.1 dollar. And 30 workers are allocated for each translation evaluation.

Finally, this experiment requires a questionnaire to each workers, and collected the information about workers in this task. The questionnaire includes the nationality, living place, and mother language of worker. These results are explained in Chapter 4 The questionnaire also have the comments by free-text. Most workers, however, left no comment, if exists, there are meaningless comment such as "No." This result suggests the crowdsourcing workers are lazy, but performed the forced tasks.

3.3 Screening Workers

In crowdsourcing, we have to consider the spammers for quality management. Spammers are the workers that are unwilling to answer honestly, and that submit a wrong answer. They includes worker with less ability or just program.

This research applied two method to remove spammers: One is the access limitation of the country the network connects, the other is the screening test.

As the former, connection from US is only allowed to perform the evaluation task. This is because the related work showed the 80% of workers are from US or India, and the worker should know both Chinese and English. Workers from China are not so large, so the limitation for quality can improve the quality and less decrease the number of worker. This type of worker screening is performed by access limitation that is inborn with MTurk This is based on IP address shutdown. Preliminary test showed workers from except US submit very low-quality result (Example: all evaluation has 5 score), we determined to accept workers only from US.

The latter, screening test, is the ability test requesters can make and impose to worker-candidate. If a worker want to perform this task, they should pass this screening test. Platform of screening test is prepared in MTurk, so we designed the screening test and the interface for qualification test. Sometimes screening test is imposed to certificate the worker is not robot spam, This experiments, in contrast, checks the ability of translation.

Figure 4 shows the screening test we imposed to workers. The test consists of the question to select the correct English translation of very easy Chinese sentence from the four candidates. The worker willing to perform this task

Figure 4: Screening Test for Chinese to English Translation Ability

should answer two questions correctly. In this screening test, the Chinese source sentences are "I'm very good at Chinese." and "What's the date today?" These source sentences are provided in a form of image, not text. This is because some spammers may use machine translation to pass the exam. It can be used only in languages whose typing is difficult for workers. Screening test checks the Chinese to English translation ability, because the experiment is Chinese to English translation evaluation. The test is designed very easy for basic learner, that focused on suppressing the worker decrease. In preliminary experiment, the difficult test is imposed. It showed the very less collection of workers. Especially, in free-text task, the result of screening test cannot be checked automatically. This requires human assessment, that costs very high for requesters. From this reason, we prepared the selection test that can easily

checked. However, there are trade-off relation between worker's quality and crowdsourcing collection efficiency. Further survey is required for obtaining optimal test. This experiment requests easier task than translate, so this easy screening test has showed the effect to a certain degree.

3.4 Implementation of Experiment

The experiment environment was implemented by Roughly, the process of the implementation is as below:

1. Create qualification test
2. Create the template for the HIT
   (a) Set the task description
   (b) Create the design of task
3. Upload CSV and request the task

Firstly, in 1, we can create the qualification test. In most cases, qualifications such as access area limitation can be used only by checking the box in 2a. To design the original qualification test, access via API is required. Then we created Qualification test written in HTML. Also, the answer and score of the test should be defined. We have to use the XML file for scoring the result of test. After creating the qualification test, qualification ID comes out.

Secondly, we create the HIT template in 2. This process seperated in two parts: one is to describe the HIT information and the other is to design the interface of the HIT. The description includes title, the qualification of workers, reward per one HIT, and number of assignment per one HIT. In this experiment, title is "Evaluate Chinese-to-English translation", for example. Also, qualification is set as "living in US" and "Chinese-English translation test". After then, we have to design the interface based on HTML. In this experiment, we implemented only by HTML. HTML, the variable of CSV is required. The variable is defined by the form of `${name}`, this name can be used in 3. Some tasks related to multimedia uses complex APIs for impose the HITs to workers.

Finally, upload CSV and request the task. CSV should have all the column name of variables in 2b. Each line in CSV file determine the each HIT, so the number of lines of CSV file determine the total cost for CSV request. Uploading

the CSV file, and then MTurk shows the confirmation screen. If there are no problems, clicking the button can request the HIT.

In this way, we created the 5-range translation evaluation task on crowdsourcing. The result and analysis of the crowdsourcing evaluation data are explained in the next chapter.

Chapter 4 Analysis of Crowdsourcing Evaluation

This research analyzed the crowdsourcing evaluation from those points of view:

- The quality of each worker's evaluation
- Improvement of evaluation quality by number of workers
- The difference by aggregation method of multiple evaluation
- The number of workers with time

This chapter firstly explain the comparative criteria between crowdsourcing and professional evaluation. After that, we show the result of analysis from each viewpoints.

4.1 Comparative Criteria

We measured the professional and crowdsourcing by correlation coefficient, MAE (mean absolute error), and VAE (Variance of Absolute Error) , MRE (Mean relative error) et, al exists as other metrics. But our evaluation score is ranged 1-5 so there are no difference between MAE and MRE.

We explain each evaluation standard below: In all the equations below, x, y are the variable vector, and n is the size of vector. In this research, x and y are the evaluation vector for each translation whose element is the score for a translation.

Correlation Coefficient Correlation coefficient (Pearson product-moment correlation coefficient) is the measure of linear similarity between two variables. Definition is below:

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

The range of correlation coefficient is -1 to 1, and is the absolute score of correlation is large, two variables are like linear relation. It should be noted that the correlation coefficient measures relative similarity, and it doesn't reflect the absolute difference of evaluation score. To solve this problem,

we have to use MAE and VAE at the same time.

MAE MAE (Mean absolute error) refers how close the two variables. the definition is the equation below:

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \tag{2}$$

MAE is 0 or more, the lower MAE means the closeness of the two variables. MAE can be estimate the expectation of the difference of each element between two evaluation vectors.

VAE VAE(Variance of Absolute Error) is the variance of absolute error.

$$\frac{1}{n}\sum_{i=1}^{n}(|x_i - y_i| - MAE)^2 \tag{3}$$

Even if MAE is low, there are some evaluation of big difference when VAE is high. So we have to measure both MAE and VAE for the analysis.

4.2 Summary of Result

Firstly, we show the statistics of workers obtained from questionnaire. 47 workers are collected in this experiment, and the result of questionnaire is as Table 3 and Table 4. As for the country they live was asked in the questionnaire, it has become the United States all as described in the section on screening.

Result indicates most of workers are from China and the native language is Chinese. As for rewards of workers, average hourly rate was 7.8 dollar. This is close to the guideline of Amazon Mechanical Turk (about 8 dollars, equals to the part-time job in USA).

Table 3: Mother Language of Workers

Chinese	English	Indonesian
43	3	1

Table 4: Country of Workers in Born

China	Taiwan	USA	Singapore	Indonesia
32	7	5	2	1

4.3 Analysis of Evaluation

4.3.1 Analysis of Individual Evaluation

As the analysis of crowdsourcing evaluation, we show the individual evaluation result and comparison with professional evaluation. If the difference is very small between individual evaluation score by crowdsourcing and professional score, all the individual results by crowdsourcing can replace the professional evaluation.

Figure 6 and figure 5 indicate the difference between individual evaluation and professional evaluation.

Summary is calculated by the difference between individual score of professional and crowdsourcing evaluation. Not all workers are evaluated all the translation, so we cant measure the translation evaluation. So we used the absolute error between individual evaluation and the professional evaluation to that translation.

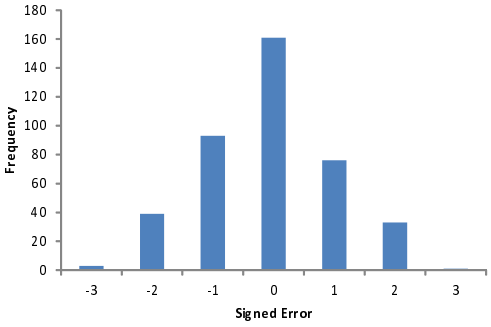


Figure 5: Histgram of Individual Signed Error

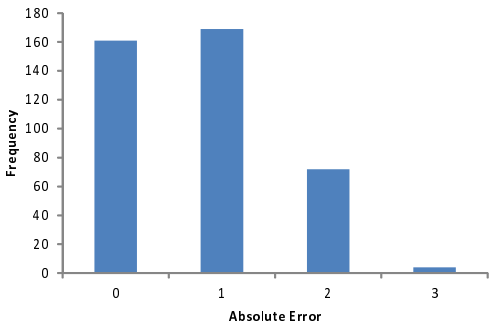


Figure 6: Histgram of Individual Absolute Error

Figure 5 represents the signed difference and figure 6 indicates the absolute difference. From Figure 5, average of error is calculated as -0.08, while its variance is 1.2. If all the crowdsourcing evaluation is averaged, there are almost

no difference from professional workers. Distribution is almost symmetrical, which shows it is close to the normal distribution.

As for Figure 6, the absolute error of most evaluations is within two, whose distribution is approximately of $0 : 1 : 2 = 4 : 4 : 2$ The average absolute error is 0.8, the variance was 0.57. This indicates that the evaluation results are different from those made by professionals often compared to the evaluation by experts. These results suggest it is difficult to use score by only one worker as an evaluation value of the evaluation evaluation.

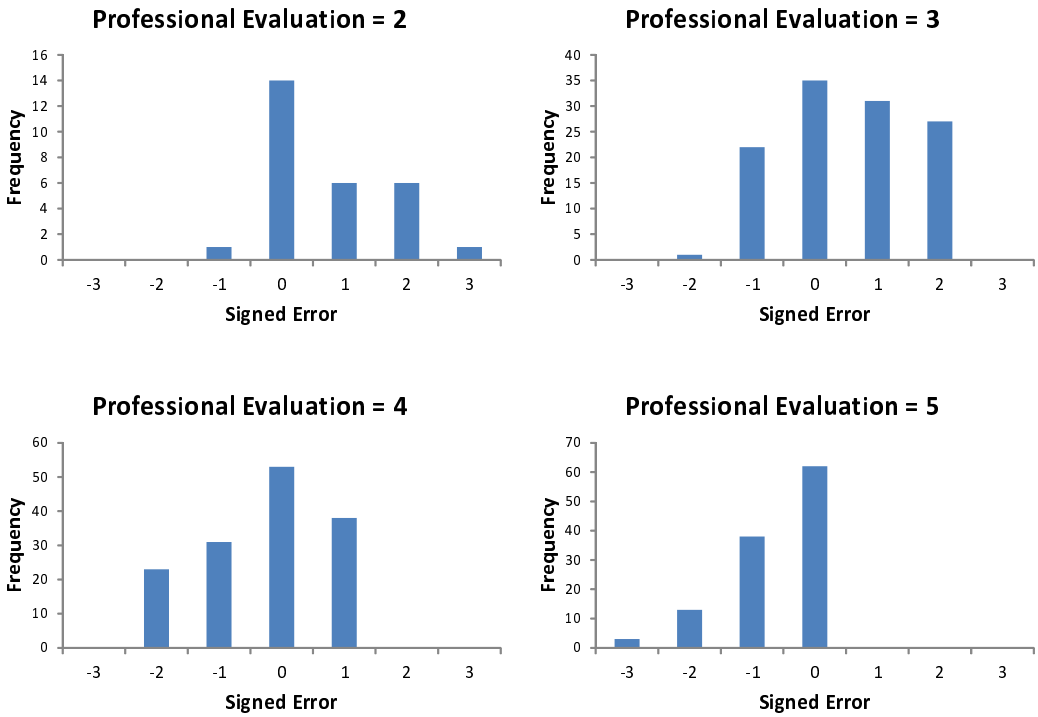


Figure 7: Histogram of Each Score by Professional

Next, we have investigated whether there are differences in the evaluation of non-professionals by the value of expert evaluation. In this experiment, distribution of error may not be symmetrical according to the professional evaluation even if the whole evaluation is normal distribution. This is because the evaluation value is in the range of one to five. If professional evaluation is four, error by crowdsourcing never becomes more than 1, for example. Also, if professional

evaluation is two, the error is less than or equal to -2.

Figure 7 is a set of histogram that shows the signed error distribution with professional score. Each histogram represents the professional evaluation of two to five. If professional evaluation was two or three (upper left and upper right in Figure 7), it seems that crowdsourcing evaluation is higher than the professional evaluation in many cases. In particular, in the case of the professional evaluation of 2, cases with a worse score by crowdsourcing than expert is less than or equal to 10% , the average signed error was 0.71. However, it could not be affirmed because there was only one translation is scored two by professional.

Conversely, when professional evaluation is four or five (lower left and lower right in Figure 7), there is a tendency to put a slightly lower than professional. In the case of score five of professional evaluation, although it is unavoidable because there is no higher evaluation from experts, the average error was the result of -0.62. We didn't prepare the translation with one score of the professional evaluation, so there is no distribution graph. If exists, estimated trend is also reflected in more extreme than in the case of the score two.

From the distribution of errors for each evaluation value, it is difficult to simply use crowdsourcing evaluation by one worker for each translation.

Therefore, it is necessary to improve the quality of the results by aggregating multiple evaluation result.

4.3.2 Analysis of Aggregated Evaluation

We performed an analysis of quality improvement by crowdsourcing translation evaluation by aggregating the work of more than one person. Evaluation results of more than one person was summarized by three aggregation methods of average, median, and mode value for each translation. In this study we measured and analyzed difference of each aggregated score and the expert evaluation. In general, the amount of information that will reduce the amount of information that remains after the aggregate has the largest average value, and decreased by the median, and the mode.

Table 5 shows the summarized analysis results of the evaluation of all workers. Calculated correlation coefficient, MAE, and VAE for each translation. As a result, the error of crowdsourcing evaluation and professional was most

Table 5: Comparison of Aggregated Score to Professional Evaluation

Aggregation	Correlation Coefficient	MAE	VAE
Mode	0.58	0.5	0.68
Median	0.81	0.29	0.20
Average	0.78	0.53	0.08

small by using median. However, for VAE, resulted average value is very small. This is caused by the real number of average. Median and mode can have only integer number among one and five. But average can have any real number.

However, we can not deny the possibility it has been evaluated by chance is close to the median for an expert evaluation and the most current data set from only this result. So using the experimental results of this study, we run a simulation to estimate the error in the evaluation of aggregate results and professional result. The method is as follows.

Algorithm 1 create a virtual environment of N evaluations. Then, we calculate evaluation score in each environment. For 1000 times creation of evaluation environment, the population and evaluation of human crowdsourcing, sampled 1000 times of size N allowing repetition. The output of the procedure MAE_j is the comparison result for each sampling. For all 1000 sets of evaluation, compare average, median, and mode score with professional evaluation to obtain MAE and variance of MAE. In this simulation, we do not consider the workers were evaluated more than one translations. That is, the number of people at random sentence extraction for the evaluation of all translation. Executed this algorithm from $N = 1$ to $N = 29$ to analyze the change of MAE. There is assumption that obtained data is representative of the population.

Figure 8 represents the change of the MAE by increase of the workers. Average, Median, and Mode in the figure means each aggregation method. When using Average and Median, it can be found that the MAE decreases with number of people, the translation evaluation score is improved. Further, it can be seen that the average MAE gets smallest in the case of using the median. However, when using the mode value, translation evaluation value was not much

Algorithm 1 Simulation Procedure

Input: Professional evaluation P and crowdsourcing evaluation S, number of evaluation N

Output: Average Mean Absolute Error MAE by each aggregation

N /* Number of evaluations for each translation */
T /* Size of evaluated translations */
$S = \{s_1, ..., s_T\}$ /* Set of evaluation by crowdsourcing.
each s_i has N evaluation scores */
$P = \{p_1, ..., p_T\}$ /* Set of professional evaluation */
p_i /* Professional score by 5-range evaluation */
$test_{ij}$ /* The evaluation set produced in this procedure
$test$ is empty set at the initial state */
$MAE \leftarrow$**null**
for $j = 1$ to 1000 **do**
    for all s_i **in** S **do**
        $test_{i,j} \leftarrow sample(s_i, N)$ /* $sample$ is a function that extracts i elements from s_j allowing repetition */
    end for
    $MAE_{j,average} \leftarrow average(test_{,j}) - p_i$
    $MAE_{j,median} \leftarrow median(test_{,j}) - p_i$
    $MAE_{j,mode} \leftarrow mode(test_{,j}) - p_i$
end for
return

improvement. The problem in using mode is caused by the unstableness of the evaluation sentence difficult. For example, for a sentence of three, if evaluation of 30 people was $\{2, 3, 4\} = \{10, 9, 11\}$, the evaluation of experts specializing in the evaluation value makes difference when using the mode. Considered a result of these cases occurred in more than one sentence, and the evaluation of the value of the most frequent ones were bad.

As the variance of MAE, score aggregated by median resulted in highest variance. In particular, the median gets high when the number of people is

even, so the evaluation graph has become jagged shape. This represents that the translation vary greatly, and the median has resulted in much lack stability evaluation. There is also the definition of the median of the evaluation if the number is an even number as one of the reasons. Because it is median of {1,2 } is calculated as 1.5, the difference at least 0.5 would come out as compared to professionals.

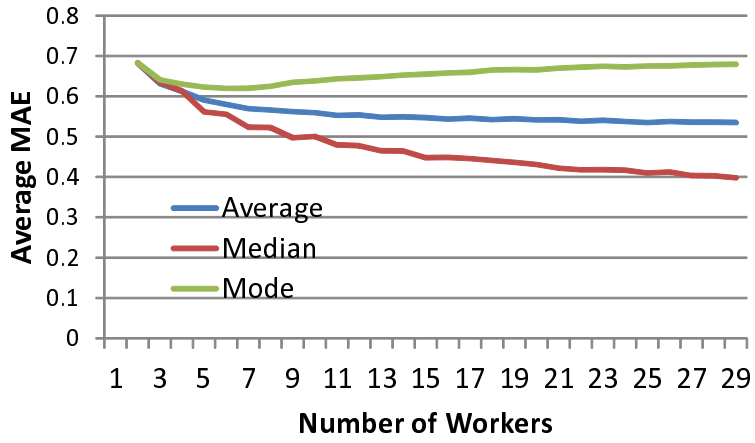


Figure 8: The Change of MAE by Number of Workers

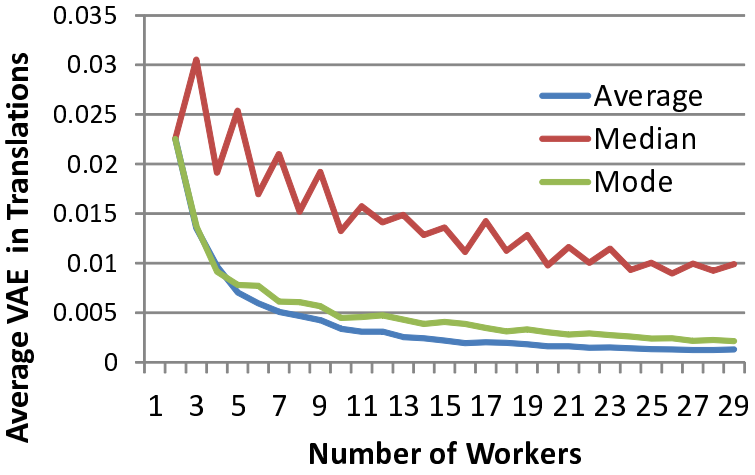


Figure 9: The Change of Variance of MAE by Number of Workers

4.3.3 Analysis of Each Translation

In addition to the analysis of the overall assessment, we performed a more detailed analysis on the translation of the individual. Here, we have investigated two translations and its evaluation by both crowdsourcing and professional. 6 is the source sentence and translation for analysis. These sentences are selected by the score of professional evaluation. Due to the 4.3.1, there should be difference by the score. We found both similarity and difference by professional score from two histograms.

Table 6: Concrete Example of Translation

ID	Chinese	English	score
01	京报网讯（记者陈嘉堃）今晚，CBA即将展开常规赛半程最后一轮角逐，北京金隅队坐镇主场迎战浙江广厦队。	Jingbao network news (correspondent Chen Jia) tonight, CBA is about to launch a regular half last round rivalry, bbmg against Zhejiang guangsha head the home teams.	3
03	虽然距离中国足协规定的内、外援转会窗口开启日期还有一段时间，但多支中超劲旅选择早早动手，频频挥舞着支票四处“抢人”，通过内引外联扩充球队实力。	Although it is still some time before the official transfer period set by the Chinese Football Association arrives, various Chinese League clubs have already begun making offers of large amounts of transfer fee in a bid to attract both local and foreign talents to strengthen their club's prowess.	5

Table 7 and Table 8 are histograms of the distribution of crowdsourcing evaluation used as the evaluation value of professional was five points, and

Table 7: Evaluation Histogram of Sentence ID 1

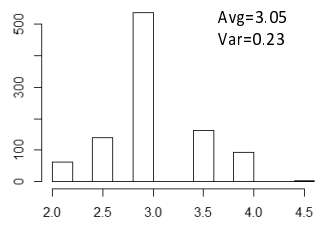
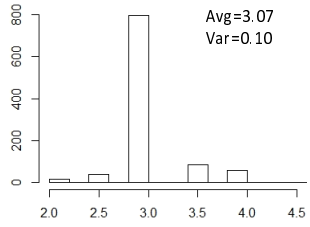
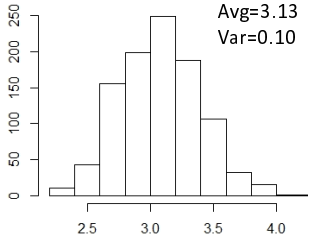
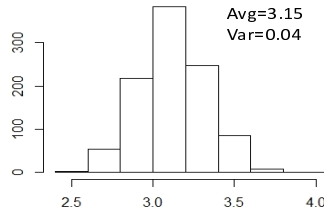
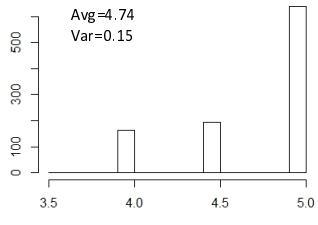
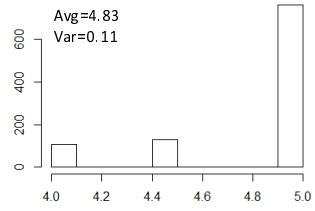
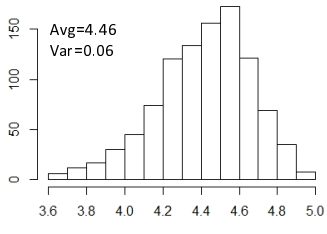
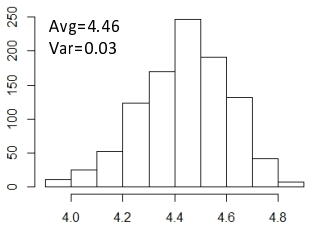
Aggregation	Number of Workers: 10	Number of Workers: 20
Median		
Average		

Table 8: Evaluation Histogram of Sentence ID 3

Aggregation	Number of Workers: 10	Number of Workers: 20
Median		
Average		

three points. These histograms are created by Algorithm 1. In these figures, ave means the average of 1000 times run score. Also, var means the variance of 1000 times run. Each histogram shows the 1000 times run by four pattern of worker size of 5, 10, 15, 20, which shows the evaluation score and frequency. That is, regarding this experiment data as the mother set, the distribution is analyzed for two aggregation method. In addition, each histogram has the average score and variance of 1000 times run.

Firstly, we can see the improvement by increase of workers. From both figures, the variances of distribution get decrease by the number of workers. It indicates there are more correct workers than incorrect workers in the field of evaluation of translation. From histograms, the frequency of correct evaluation is getting larger by increase of workers.

Secondly, what is seen from these figures, is that median is better when the professional evaluation was five, whereas average value is better when the professional evaluation was three. Median and the variance value have shown little error is not related to the value of expert evaluation, in the case of using the average value of three points better error decreases. Therefore, from the viewpoint of better variance, it is considered of the average value to be a value close to the professional evaluation. As a result, when using the crowdsourcing evaluation, changing the aggregation method according to the average value to obtain a good evaluation. We're going to discuss the implementation in Chapter 5 in detail.

4.4 Analysis about Complete Time

Finally we analyzed the turn-around time from task request to task finish. Here we call complete time.

Figure 10 is a diagram of the relation between number of workers and date elapsed since the day task was requested. Also, the diagram contains the preliminary experiment requested by half of the reward. It can be seen that the number of workers decreases each time from the start. These workers performed the most of the requested task, so there are no difference between task performed and worker. Although it's risky to determine from only these results

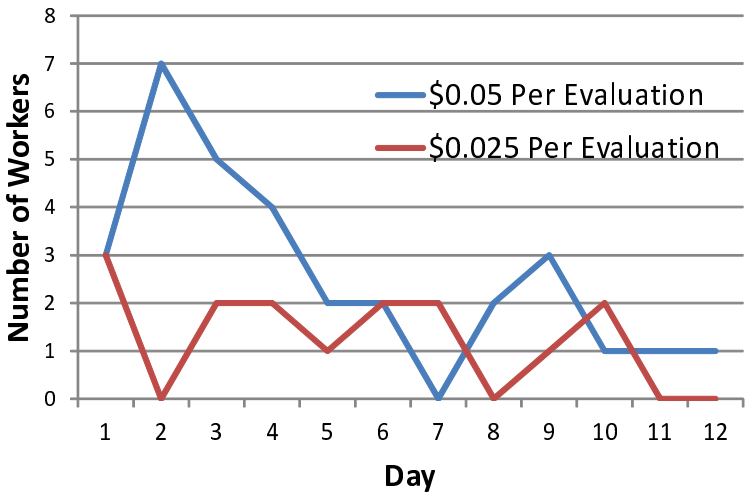


Figure 10: Graph of Process Dates and Task Result

due to the small size of sample, number of new workers will decreased.

Two reasons are listed for this result. One is because as the task is no longer the latest, disappeared from the search results. In particular, those who will do any task perform a task randomly. So when the task disappears from the top search results, the workers that found our task decreases. In MTurk, due to the default search setting "Hits Creation Date (Newest first)" for most workers, it becomes a task which can not be seen [12]. To solve this problem some can be proposed: to design the more useful interface of the platform, to ask the task in another crowdsourcing platform, or, more simply, to register a new task to keep a "fresh" state of the task.

The other reason is the limited worker with translation ability to access to MTurk. Daily workers access MTurk every day, and perform immediately if they are feasible to complete the task. Therefore, it is considered that the workers capable of translation that have not finished work yet has gradually decreased with time. In particular, only 22 of the search results when searched by word "Translation" in MTurk, and unlikely to be affected by the above. There are not many workers in MTurk who purposely going to perform the translation.

In addition, the relationship between the number of worker and the reward correlation appeared clearly. With the cheap reward, there are very few workers;

less than one worker per hour carry out the translation evaluation task. This may be caused by the effect of the relatively low cost. Among the preliminary experiment, only this task was set as the $ 0.025 per translation.

4.5 Summary of Analysis

The summary of analysis is below:

1. Crowdsourcing evaluation of translation has some deviation compared to the evaluation by professionals
2. Evaluation value is approaching to professional as increase of workers, but the change of the evaluation value for the number the slower
3. With respect to aggregation method of the results of more than one person, if the evaluation value very good / bad evaluation, median value is better; if the score is close to the middle, average can obtain the score with minimal difference from professional evaluation
4. Increase of time of the task-complete has become like exponential distribution, the worker will decay over time dramatically
5. For the increase of the cost, the positive effect is observed, while the increase of the time the effect is very small

For 1 and 2, they should be a result of the course. However, important finding is that the workers with sufficient quality gather by applying the appropriate screening test, using crowdsourcing. In addition, the point shows the effect of the accumulation of collective intelligence by crowdsourcing is valuable.

Secondly, better aggregate method can suppress the impact of incorrect results of the evaluation in crowdsourcing evaluation by 3. If the expert evaluation is 5, the worker can not exist to evaluate higher than 5. More than half of workers evaluated correct result 5, and have resulted in a median is calculated to 5. Further, with respect to the translation with professional evaluation of three, in the middle of evaluation criteria, the evaluation distribution by crowd is results of the worker close to the binomial distribution, which affect the average is lowered as compared to professional score of five.

Finally, from 4 and 5, the most working period is important the first few days since requested the task, the long-term effect of extending the time period

it can be seen that the low. This suggests that workers who examined daily translation evaluation tasks carry out most of these evaluations. During evaluation of experiment, workers with same ID appear in most experiments including preliminary task. They performed all the task as soon as the task is registered. That is, in the long term, we have to coexist with them.

These results suggest that, to control the quality and cost, following guidelines for the evaluation of the translation is required.

1. Imposing screening test in accordance with the language pair you are trying to evaluate
2. If you want to increase the number of workers, it is necessary to set a high reward
3. Even if you are concerned about the complete time, that we need about one or two days.
4. You have to change aggregation method to obtain optimal score from crowdsourcing

Chapter 5 Crowdsourcing System for Translation Evaluation

Analysis showed the sufficient number of workers can be no comparable scores of professional evaluation value taking the median or mean value. However, it is necessary to take a long period of time and cost more to collect a lot of result for the reduction of the MAE. There is a trade-off among cost, time and quality for the requester, the requester himself have decided by fumbling it can not be said to be a practical crowdsourcing translation evaluation.

Therefore, based on the findings of in Chapter 4, we propose an interface to request the task to Amazon Mechanical Turk. The interface has a function based on an analysis of the past, as small as possible the error between the expert. It also manages previous work, and indicates the quality evaluation of how much has been achieved so far. The evaluation of multiple workers are based on the analysis result.

5.1 Aggregation Method of Crowdsourcing Evaluation

We can expect to be able to obtain better evaluation close to professional by changing the aggregation according to the average of crowdsourcing evaluation, as the result of analysis. In the implementation of this study, we propose an algorithm to select the aggregation method, the analysis of multi evaluation of crowdsourcing.

2 is an algorithm used to aggregate the evaluation by crowdsourcing in the implementation. This algorithm firstly analysis to obtain an average value by crowdsourcing, using the median or average value is used depending on whether the average value is close to the middle of evaluation. By using average value if it is between 4-2 mean value, by using median otherwise we aim to evaluate more closer to the professional.

5.2 Implement of Crowdsourcing Evaluation System

In the following, we describe the translation evaluation system in detail. This system is composed of two modules: task request module and task management

Algorithm 2 Aggregation Algorithm

Input: crowdsourcing evaluation for one sentence S

Output: Aggregated crowdsourcing score s

1: $S = \{s_1, ..., s_N\}$ /* Set of evaluation by crowdsourcing. */
2: **if** $2 < average(S) < 4$ **then**
3: **return** $average(S)$
4: **else**
5: **return** $median(S)$
6: **end if**

module. This implementation, enables the requester to predict whether results will be obtained and their quality. If the prediction is out, it quickly fixes from the management module is possible.

5.2.1 Amazon Mechanical Turk API

As usage of Amazon Mechanical Turk, it's available to request and retrieve the results of the task not only from the Web directly, but through the API offered by Amazon. API can be accessed in REST or SOAP format is possible, and is available to anyone who have a account. In this work, through REST format, we have designed a translation evaluation system via crowdsourcing to retrieve data from the Amazon Mechanical Turk.

Here's the difference between a task through API and through the Web page to request a task in Amazon Mechanical Turk. Design of the task is assumed that the advance is performed on the Web. Figure 11 is a scene to upload a CSV file through the interface of Amazon Mechanical Turk to start a translation evaluation task. For the task template is already registered via Web, you can run the task if the task to check CSV upload and task requests from the screen after login is possible.

The same function as Figure 11 is made through GET messaging of HTTP in Figure 12. Sending message to Amazon Mechanical Turk API[1)], it is possible to request the same tasks. The original HTTP message is encoded in Chinese and English sentences, but is decoded for lucidity. Here, one that is given when

[1)] https://mechanicalturk.sandbox.amazonaws.com/

you create the interface in the Web, HITLayoutId is set translation evaluation tasks to be performed in advance. In addition, AWSAccessKeyId and Signature are represented by the 20-digit and 40-digit code. They are required for security authentication.

In this way, it is possible to request the Web or task via Web API so that the requester can easily ask workers to evaluate translation. Further, although it is easy to request a Web page from a different task MTurk. The only thing to be careful is to pay close attention to the user authentication.

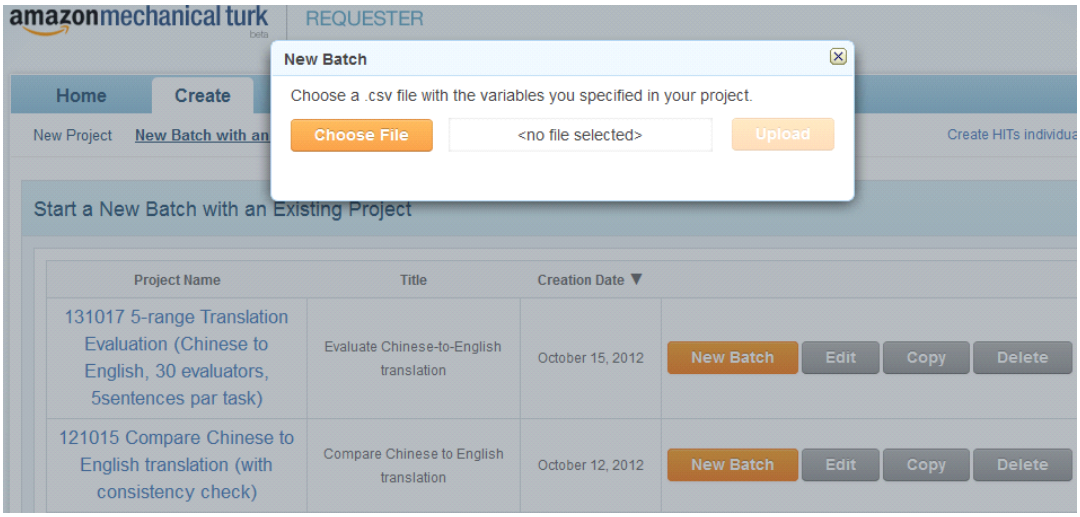


Figure 11: Task Request Via Amazon Mechanical Turk Web Page

5.2.2 Task Request Module

Using Amazon Mechanical API, we designed the crowdsourcing evaluation system of translation. Figure 13 shows a system structure of the crowdsourcing evaluation system.

Here, we will explain the task request module, upper side of Figure 13. First, a requester specify the language pair, source sentence, and translation sentence for the task. Source text and translation, can be written in the form of CSV, it is possible to describe any number of translation. The requester fix either time or cost for crowdsourcing evaluating, then the system shows the estimated cost and time. If there is a translation system the requester wants to obtain the evaluation score within a week, he inputs a The system searches

```
AWSAccessKeyId=/* MTurk Key ID*/
Operation=CreateHIT
Service=AWSMechanicalTurkRequester
Signature=/* MTurk signature */
Timestamp=2013-02-01T06:22:01Z
Version=2012-03-25
Reward.1.Amount=0.02
Reward.1.CurrencyCode=USD
HITLayoutId=/* Layout ID*/
HITLayoutParameter.1.Name=Chinese
HITLayoutParameter.1.Value=  我可以说汉语
    HITLayoutParameter.2.Name=English
HITLayoutParameter.2.Value=I can speak Chinese
```

Figure 12: HTTP Messaging to Request via API

the database of past evaluation based on the language pair, and then output the estimated cost and quality within a week. Evaluation quality is output in the form of MAE. For example, the text output from the system is as follows. "The predicted value is about 0.3 $\pm$ evaluation of professional evaluation value and will be obtained after 7 days" Looking at the output of the system, the requester makes a determination of whether or not to use the crowdsourcing eventually.

If a requester decide to make use of the system, system make use of the API of Amazon Mechanical Turk, based on template made previously that corresponds to the language pairs. After that, the module registers the task ID API returns, connecting with the user ID. Note that it has become a form of the task is created one by one for each translation. In addition, the screening tests is imposed to the worker as described 3.3 at this time. Task per one translation is 30 by default setting, but it can be changed by the requester to specify the maximum number possible.

As mentioned in 2.2, task completion time of crowdsourcing has been an

exponential distribution, so predictions about task completion time is difficult. However, when performing the task the same, the result of the number of workers gathering range that is the same. By storing the evaluation data of the past, evaluation of the number of workers to some extent.

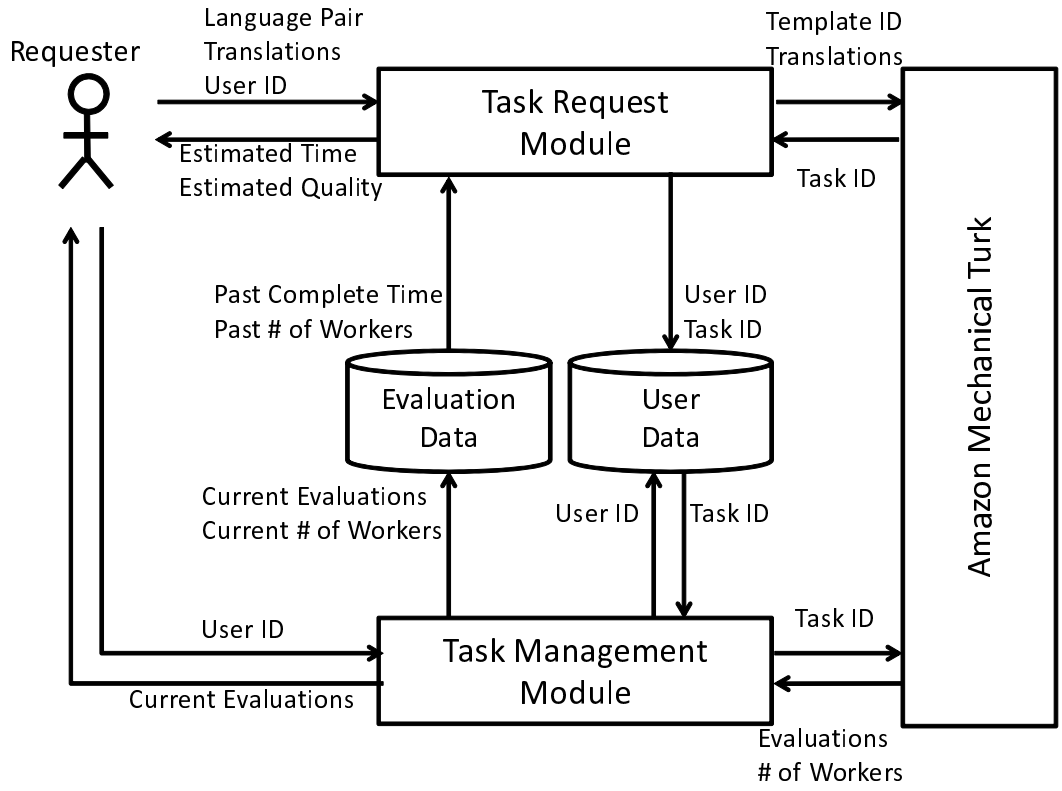


Figure 13: System Structure of Requester Interface

5.2.3 Task Management Module

The other module, task management module, enables requesters to manage of the execution status of the task by the task request module.

Task management module is the lower side of Figure 13. First, a requester logs in with the user ID of themselves. Then, task management module to get the current value of the evaluation, the evaluation of the current to get to it. Indicator of whether the current evaluation and how close to the professional evaluation value is given by the variance of the number of workers. What is considered a good indicator, and close enough to the expert evaluation, the

status is displayed in blue. On the contrary, evaluation with not good status indicator is displayed in yellow. Depending on the completion status of the task, the request is then may be terminated in the middle of the evaluation, it is also possible to extend the time period vice versa. If you abort the evaluation, the summary of the current evaluation value is obtained as a result.

Evaluation data is stored in the database anonymously. Data can be performed with high accuracy by using the relationship between the elapsed time and the number of tasks for each language pair, newly updated, to guess the number of workers for another requester is possible.

Chapter 6 Discussion

Here, we will discuss how crowdsourcing evaluation about what we will achieve, based on the analysis and implementation so far. In particular, we consider the more practical approach to the translation evaluation using crowdsourcing and future comparison of existing crowdsourcing evaluation method.

6.1 Comparison Crowdsourcing and Other Evaluation Methods

The following is the difference among the translation evaluation method. In particular, the time, quality, and cost that has been discussed in this work.

Table 9: Comparison among the translation evaluation methods

	Automatic	Professional	Crowdsourcing
Cost	Low	Very High	Low - Medium
Time	Fast	Slow	Medium - Slow
Quality	Low	Very High	High
Note	Requires reference translation	Requires training when evaluated by multiple professionals	

As seen from the table 6.1, time, speed, there is a trade-off between cost, respectively.

Crowdsourcing translation evaluation method can perform a evaluation that could not be met by existing evaluation scheme. In particular, crowdsourcing has great advantage in large scale evaluation due to the scalability. Also, as mentioned in Chapter 5, crowdsourcing evaluation of translation is flexible by adjusting the reward and working conditions by period, depending on the purpose of the request. In the table, crowdsourcing evaluation has wide range for each metric. An evaluation requester can change the condition for evaluation to determine what they put in priority.

Also, here is the detail of the contents of the column Note. As mentioned

in Chapter 2, automatic evaluation method for translation requires reference translation to improve the evaluation quality. In most cases, bilingual corpus have more than one translations for each source sentence. Multiple reference can improve the quality. As for professional evaluation, a requester often contact to many bilingual professionals for massive evaluation. There occurs the problem of the different standard by the professionals. So they have to train using prepared translation with fixed evaluation score. In the training, they have to come to the agreement of each sentence. These overhead costs both in money and speed.

In spite of these problems in existing translation evaluation method, crowdsourcing doesn't need high-cost preparation before asking the evaluation task. For example, instruction is common with each evaluation standard. We don't have to write them twice. Also, screening test can be easily implemented with language knowledge of elementary level. In fact, crowdsourcing is fast and cheap to start and very flexible even after requesting evaluation.

In order to take advantage of crowdsourcing evaluation of translation, it is considered to be useful in the following cases.

- Translation service is updated every day, and changing in translation quality
- Translation service is domain-specific, we can not afford the cost of providing a bilingual corpus
- Translation is obtained by crowdsourcing translation, or is generated dynamically, so they can not fully guarantee the quality of the translation

In particular, the big point is that they only have to communicates with the interface of MTurk. It has good chemistry with the situation in which the translation is generated dynamically.

On the other hand, in the case of the following, there are small benefits of using a crowdsourcing translation evaluation.

- When evaluating translation for bilingual corpus statement already exists
- When the reference translation is reusable, and the cost of creating correct translation is low
- There is enough space for both cost and time, and is required for high-

quality evaluation

In other words, the evaluation translation evaluation by crowdsourcing so far is used for the case that there has not been applied to existing evaluation method. However, it is concluded that with respect to the corresponding part can be necessary to use crowdsourcing purposely not in the framework of the existing evaluation.

6.2 Further Improvement of Crowdsourcing Evaluation

Currently, crowdsourcing translation evaluation showed that although the quality is low, a meaningful evaluation of certain translation can be compared to the expert. However, since it is considered to improve by a variety of techniques, here we will discuss about the possibility of further improvement in the future.

First, crowdsourcing workers evaluated using a fluency translation accuracy, generally used in this field, for the evaluation standard, it is considered that the use of many other indicators. NTCIR9 evaluates the translation under Acceptability example. Also, NTCIR10 is going to evaluate with a basic understanding of the need for new patent. These metrics requires a knowledge of the areas of expertise in addition to knowledge of the language, so the ability of workers to meet current crowdsourcing is difficult. In the case of using a crowdsourcing, it is also possible to define new metrics due to the high flexibility. By focusing on the case of translation of the tourist is able to understand foreign tourists translation purposes is that it is possible, a new evaluation standard of the translation how easily to get to the destination, for example. It was a reliable assessment of how the evaluation was carried out in the present study the extent only is, and also how much error is evaluated by crowdsourcing than experts. Therefore, by requesting new task with new design, there is also a possibility that the optimal evaluation method for crowdsourcing born.

In addition, the discussion of the remuneration paid to the worker not been in depth. In this study, we assume that task completion time will decrease monotonically decreases with increasing compensation provisionally, the form of that function is not clear. However, study in crowdsourcing or psychology results in that the past reward is high amount, workers expect are that much

then the same amount. Therefore, it is possible to work with an operator by easily compensation may be dangerous. In particular, there is a high possibility to set different fees for the same task. In that case, crowdsourcing evaluation task with cheap reward will be never done before the high-reward ones get completed. It is necessary to study in detail about the work period when these tasks simultaneously exist in the future. Related to the problem above, there is also a problem how to implement of screening test. This is considered the effect of screening test is sufficiently obtained, it is possible that an operator with a capability is no longer perform the tasks by screening. As screening methods other than qualification test, machine-learning based method or gold-unit test is proposed so far. We have to see what occurs when these methods are applied to the task that requires profession. The analysis was carried out by comparing with evaluation score of crowd workers and professionals by the MAE and VAE. Evaluation of whether available extent as the actual crowdsourcing evaluation of translation has not been estimated. This is because we don't discuss relationships of the usefulness of the actual MAE. In order to indicate that the evaluation of translation is Feasible, it is necessary to obtain consent of the request rate, he that that the actual experiments the portion you will need. In the project of the Language Grid, there are a wide variety of tasks, such as medical document translation or translation of agricultural knowledge has been carried out, for example. In this context, I would like to get further improved by comments to the requester to conduct the evaluation crowdsourcing actually have the expertise and bilingual.

Finally, the translation evaluation in this study focused on only Chinese to English translation, we have conducted experiments using crowdsourcing in practice. Whether the same results can be obtained in other language environments is not discussed. For example, there is a possibility that in the present study because the workers are limited to residents of the United States when the target language is English, evaluation is successful. Evaluated equally for translation in English in the case where reversing the path translation in order to ensure strict evaluation, evaluation translation using crowdsourcing evaluation of whether it is useful regardless of the language want to do. In addition, it

is considered that the future will come an increase in demand even for translations between non-English languages, it is necessary to analyze crowdsourcing whether it is possible to put into practical use for the evaluation of translation in various projects .

Chapter 7 Conclusion

Various methods have been proposed until now to evaluate translation quality. However, the existing methods of evaluation have problems in terms of cost or quality, new technique was required to solve these problems. In this study, we have used a crowdsourcing-based approach for this problem to achieve a translation evaluation which is low-cost and close to professional evaluation. There is a problem in the evaluation of crowdsourcing worker alone has low reliability compared to the professional evaluation, but it can be trusted it is possible to number you have collected. Crowdsourcing workers can perform 5-range evaluation. In addition, we developed a new interface that requester determine the configuration of the evaluation. This interface has the ability to request the task and to manage the progress of the work, and the function to predict the time period for any compensation and, how much quality is obtained.

The conclusion of this research is as follows:

Feasibility of crowdsourcing evaluation of translation

Although reliability fall compared to professional evaluation, crowdsourcing evaluation of translation can be good quality when collected by a sufficient number. In particular, crowdsourcing is highly useful in terms of complete time and the cost.

Change of evaluation quality by number of crowdsourcing workers

In making the evaluation of translation quality by crowdsourcing,the effect of increase in number is very high, that the absolute error and variance of error becomes smaller with an increase in the number. Further, I have found that there is an effect that can be used to reduce the error between the median of the evaluation values very professional.

As future work of this research, the following points may be mentioned. The first one is an analysis of the work period and work quality and remuneration of workers. In this study we have set up a higher reward than a typical MTurk rewards. If the reward is reduced, there is a possibility of working efficiency falls. Therefore, by request of different tasks simultaneously compensation for the same work, it is necessary to analyze the trend of the operator. Since the data on

the remuneration of the evaluation by experts translation has not been obtained, crowdsourcing translation evaluation whether it is economically much we must investigate. The other is the demonstration of the utility of interface. Despite the implementation, in this study, an experiment in the Amazon Mechanical Turk experiments using the proposed interface is not performed. It is necessary to make sure the system is inferred evaluation results by actually moving the system whether obtained.

Acknowledgments

First, the author would like to express sincere gratitude to the supervisor, Professor Toru Ishida at Kyoto University, for his continuous guidance, valuable advices, and helpful discussions. The author would like to express his gratitude to Assistant Professor Donghui Lin, Assistant Professor Yuu Nakajima at Kyoto University for his technical advices and helpful discussions. The author would like to tender his acknowledgements to the advisers, Associate Professor Keishi Tajima at Kyoto University and Chief Researcher Yohei Murakami at National Institute of Information and Communications Technology for his valuable advices. Moreover, the author would like to express deep acknowledgements to all workers participated in the experiment at Amazon Mechanical Turk. Finally, the author would like to thank all members in Ishida and Matsubara laboratory, especially to Takuya Nishimura and for their various supports and discussions.

References

[1] Ishida, T.: *The language grid: Service-oriented collective intelligence for language resource interoperability*, Springer (2011).

[2] Goto, I., Lu, B., Chow, K. P., Sumita, E. and Tsou, B. K.: Overview of the patent machine translation task at the ntcir-9 workshop, *Proceedings of NTCIR*, Vol. 9, pp. 559–578 (2011).

[3] Koehn, P. and Monz, C.: Manual and automatic evaluation of machine translation between European languages, *Proceedings of the Workshop on Statistical Machine Translation*, StatMT '06, Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 102–121 (2006).

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

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

[6] Ipeirotis, P. G., Provost, F. and Wang, J.: Quality management on amazon mechanical turk, *Proceedings of the ACM SIGKDD workshop on human computation*, ACM, pp. 64–67 (2010).

[7] Koehn, P.: *Statistical machine translation*, Vol. 11, Cambridge University Press (2010).

[8] Paul, M., Federico, M. and Stuker, S.: Overview of the iwslt 2010 evaluation campaign, *IWSLT10: International Workshop on Spoken Language Translation*, pp. 3–27 (2010).

[9] Gamon, M., Aue, A. and Smets, M.: Sentence-level MT evaluation without reference translations: Beyond language modeling, *Proceedings of EAMT*, pp. 103–111 (2005).

[10] Paolacci, G., Chandler, J. and Ipeirotis, P.: Running experiments on amazon mechanical turk, *Judgment and Decision Making*, Vol. 5, No. 5, pp.

411–419 (2010).

[11] Ipeirotis, P.: Demographics of mechanical turk (2010).

[12] Ipeirotis, P. G.: Analyzing the amazon mechanical turk marketplace, *XRDS: Crossroads, The ACM Magazine for Students*, Vol. 17, No. 2, pp. 16–21 (2010).

[13] Chen, K. T., Chang, C. J., Wu, C. C., Chang, Y. C. and Lei, C. L.: Quadrant of euphoria: a crowdsourcing platform for QoE assessment, *Network, IEEE*, Vol. 24, No. 2, pp. 28–35 (2010).

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

[15] Bentivogli, L., Federico, M., Moretti, G. and Paul, M.: Getting Expert Quality from the Crowd for Machine Translation Evaluation, *Proceedings of the MT Summmit*, Vol. 13, pp. 521–528 (2011).

[16] LDC: L. D. A. Specification: Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations. (2002).