Crowd-based Evaluation of English and Japanese Machine Translation Quality

Michael Paul and **Eiichiro Sumita**
NICT
Hikaridai 3-5
619-0289 Kyoto, Japan
<Firstname>.<Lastname>@nict.go.jp

Abstract

This paper investigates the feasibility of using crowd-sourcing services for the human assessment of machine translation quality of English and Japanese translation tasks. Non-expert graders are hired in order to carry out a ranking-based MT evaluation of utterances taken from the domain of travel conversations. Besides a thorough analysis of the obtained non-expert grading results, data quality control mechanisms including "locale qualification", "on-the-fly verification" and "payment" are investigated in order to increase the reliability of the crowd-based evaluation results.

1 Introduction

This paper investigates the feasibility of using crowd-sourcing services for the human assessment of translation quality of non-English target languages. Shared evaluation tasks such as WMT (Callison-Burch, 2009) and IWSLT (Bentivogli et al., 2011) carried out crowd-based evaluations for English translations and reported moderate agreement rates between non-expert and expert graders. In contrast, this paper focuses on the crowd-based evaluation of translation tasks having Japanese as the target language.

The MT evaluation experiments were carried out using utterances taken from the domain of travel conversations. The translation quality of the MT engines was evaluated using (1) the automatic evaluation metric BLEU (Papineni et al., 2002) and (2) human assessment of MT quality based on the *Ranking* metric (Callison-Burch et al., 2007). Non-expert graders were hired through the CrowdFlower[1] interface to Amazon's Mechanical Turk[2] in order to carry out the ranking-based MT evaluation. Besides a thorough analysis of the collected grading results, we also investigate different data quality control mechanisms in order to increase the reliability of crowd-based evaluation results.

The experiments carried out in this paper revealed that high-quality evaluation results can be collected even for non-English languages given that control mechanism carefully tailored to the evaluation task at hand are in place.

[1]http://crowdflower.com
[2]http://www.mturk.com

2 Mechanical Turk Demographics

Past surveys on the demographics on MTurk users indicated that most of the workers come from the US and India. An examination of MTurk workers carried out in (Ipeirotis, 2010) reported contributions of 468 workers from the US, but only 2 Japanese out of 1000 MTurk workers. This analysis indicates that not many native speakers of Japanese are to be expected for the MT evaluation task described in Section 3.

3 MT Evaluation Task

The crowd-based MT evaluation experiments are carried out using the multilingual *Basic Travel Expressions Corpus* (BTEC), which is a collection of 160k sentences that bilingual travel experts consider useful for people going to or coming from another country (Kikui et al., 2006). The parallel text corpus was randomly split into three subsets for evaluation purposes (*eval*, 250 sen), the tuning of model weights (*dev*, 1k sen) and the training of MT engines (*train*, 160k sen). Furthermore, three subsets of varying size (80k, 20k, and 10k sentences) were randomly extracted from the training corpus and used to train four statistical MT (SMT) engines on the respective training data sets.

The translation results evaluated in this paper were obtained using fairly typical phrase-based SMT engines. For the training of the SMT models, standard word alignment and language modeling tools were used. Minimum error rate training (MERT) was used to tune the decoder's parameters. For the translation tasks having Japanese as the source (SRC) language and English as the target (TRG) language and vice versa, an in-house multi-stack phrase-based decoder was used.

Table 1 summarizes the translation quality of the SMT engines according to the standard automatic evaluation mertic BLEU (Papineni et al., 2002). Scores range between 0 (worst) and 1 (best).

Table 1: Translation Quality (BLEU)

Language		MT Engine			
SRC	TRG	160k	80k	20k	10k
en	ja	0.2858	0.2538	0.2100	0.1941
ja	en	0.2447	0.1995	0.1535	0.1257

Human assessments of translation quality were carried out using the *Ranking* metrics where graders were asked to *"rank each whole sentence translation from Best to Worst relative to the other choices (ties are allowed)"* (Callison-Burch et al., 2007). The *Ranking*

evaluation was carried out using a web-browser interface and graders had to order four system outputs by assigning a grade between 1 (*best*) and 4 (*worse*).

The most informative indicator of the quality of an evaluation dataset is given by the agreement rate, or grading consistency, both between different judges and within the same judge. To this purpose, inter- and intra-annotator agreement between MTurk workers and expert graders was calculated using the *Fleiss' kappa coefficient* κ (Landis and Koch, 1977):

$$\kappa = \frac{\Pr(a)-\Pr(e)}{1-\Pr(e)},$$

where Pr(a) is the observed agreement among graders, and Pr(e) is the hypothetical probability of chance agreement. (Landis and Koch, 1977) distinguishes the following six levels of grader agreement: "*none*" κ <0, "*slight*" κ $\leq$0.2, "*fair*" κ $\leq$0.4, "*moderate*" κ $\leq$0.6, "*substantial*" κ $\leq$0.8, and "*almost perfect*" κ $\leq$1.0.

4 Crowd-based MT Evaluation

To counter the high costs for human evaluations, crowdsourcing services such as MTurk and CF, have attracted a lot of attention as a means for collecting data at low cost. MTurk is an on-line work marketplace, where people are paid small sums of money to work on Human Intelligence Tasks (HITs), i.e. tasks that machines have hard time doing. The CF platform works across multiple crowdsourcing services, including MTurk. CF gives unrestricted access, making it possible for non US-based requesters to place HITs on MTurk.

4.1 Data Quality Control Mechanism

One of the most crucial issues to consider when collecting crowdsourced data is how to ensure quality. CF provides quality control mechanisms such as the "locale qualification" feature (to restrict workers by country) or the "on-the-fly" verification of the workers' reliability using so-called *gold units*, i.e. items with known labels, along with the other units composing the requested HIT. These control units[3] allow distinguishment between trusted workers (those who correctly replicate the gold units) and untrusted workers (those who fail the gold units). Untrusted workers are automatically blocked and not paid, and their labels are filtered out from the final data set. CF uses the workers' history to apply confidence scores ("trust level" feature) to their annotations. In order to be considered trusted in a job, workers are required to judge a minimum of four gold units and to be above an accuracy threshold of 70%.

In this paper, we investigated the dependency of the quality of the evaluation results for the following quality control features:

- *locale qualification* (LOC): restriction to official language countries; the most important control mechanism to prevent workers tainting the evaluation results.
- *on-the-fly verification* (GOLD): identification of trusted workers using control units with a known answer.
- *payment* (PAY): amount of money paid to the MTurk workers for a single HIT.

[3]The suggested amount of gold units to be provided is around 10% of the requested units.

4.2 Control Units

Control units have to be unambiguous, not too trivial, but also not too difficult. For our experiments, we selected the original corpus sentence as the main reference translation. From paraphrased reference translations[4] we selected a single reference as the *gold translation* to be included in the control units based on the following criteria: (1) it should be similar to the main reference and (2) its translation quality should be better than the best MT output of all translation hypotheses for the same input. For each paraphrased reference, we calculated the edit distance to (a) the main reference and (b) the best MT output and selected the one with minimal distance to the main reference and maximal distance to the MT output as the gold translation. The top-30 sentence IDs with the best gold translation distance scores were selected as control units for the respective translation task. For each control unit sentence ID, a random MT output was replaced in the ranking set with the gold translation. For our experiments, we distinguish two GOLD annotation schemes:

- "best-only" (GOLD^{b}): check only the gold translation, i.e., force rank '1' assignment for the best translation.
- "best+worse" (GOLD^{bw}): check the gold and the worst translation, i.e., allow rank '1' or '2' for the best and rank '3' or '4' for the worst translation.

4.3 Evaluation Interface

CF provides an *external* interface for MTurk workers to be paid and an *internal* one for which you have to prepare your own work force. The internal interface is (currently) free of charge and was used to collect judgments of in-house expert graders based on exactly the same HITs as the MTurk workers. The unit of evaluation was the *ranking set*, which is composed of a source sentence, the main reference provided as an acceptable translation, and the MT outputs of all four MT engines to be judged. The order of the MT outputs as well as the location of the gold translation was changed randomly for each ranking set to avoid any bias.

4.4 Experiment Setup

We repeated the same MT evaluation experiment using the following data quality control settings[5]:

1. *NONE*: no quality control (ja, en)
2. *GOLD*: on-the-fly only (ja, en)
3. *LOC+GOLD*: locale+on-the-fly (ja, en)
4. *LOC+GOLD+PAY*: locale+on-the-fly+payment (ja)

A HIT consists of 3 ranking sets per page and is paid $0.06 for all experiments besides LOC+GOLD+PAY where we paid 4 times that amount. In total, the evaluation costs[6] of all experiments sum up to $114 for 7 experiments, resulting in an average of $16 for the evaluation of 4 MT outputs for 300 input sentences.

[4]Up to 15 paraphrased reference translations are available for the data sets described in Section 3.

[5]IND was excluded by default for all experiments reported in this paper.

[6]The requester's payment includes a fee to MTurk of 10% of the amount paid to workers. In addition, CF takes a 33% share of the payments by the requester.

Table 2: Characteristics of Mechanical Turk Workers

Amount of Workers

	Data Quality Control Mechanism															
	LOC+GOLD^{bw}+PAY				LOC+GOLD^{bw}				GOLD^{b}				NONE			
TRG	total count	trusted *% of total*	(no overlap) *% of total*	[native] *% of total*	total count	trusted *% of total*	(no overlap) *% of total*	[native] *% of total*	total count	trusted *% of total*	(no overlap) *% of total*	[native] *% of total*	total count	trusted *% of total*	(no overlap) *% of total*	[native] *% of total*
en		–			23	73.9%	(69.5%)	[69.5%]	38	76.3%	(44.7%)	[34.2%]	8	–	(50.0%)	[50.0%]
ja	10	75.0%	(75.0%)	(20.0%)	14	71.4%	(64.2%)	[28.5%]	15	86.6%	(6.7%)	[0.0%]	10	–	(60.0%)	[10.0%]

Country of Origin

	Data Quality Control Mechanism			
TRG	LOC+GOLD^{bw}+PAY country: workers	LOC+GOLD^{bw} country: workers	GOLD^{b} country: workers	NONE country: workers
en	–	9 countries USA:15, AUS:1, CAN:1, GBR:1, MYS:1, PHL:1, BGD:1, CMR:1, SGP:1	11 countries USA:15, MKD:9, CHN:2, NLD:2, ROU:2, JPN:2, PAK:2, AUS:1, BGD:1, CMR:1, MDV:1	4 countries USA:5, AUS:1, JPN:1, MKD:1
ja	2 countries USA:8, JPN:2	2 countries USA:10, JPN:4	8 countries MKD:6, ROU:2, PAK:2, BGD:1, CHN:1, JPN:1, MDV:1, NLD:1	5 countries USA:4, JPN:2, MKD:2, PAK:1, PHL:1

Amount of Judgments

	Data Quality Control Mechanism															
	LOC+GOLD^{bw}+PAY				LOC+GOLD^{bw}				GOLD^{b}				NONE			
TRG	total count	trusted *% of total*	(no overlap) *% of total*	[native] *% of total*	total count	trusted *% of total*	(no overlap) *% of total*	[native] *% of total*	total count	trusted *% of total*	(no overlap) *% of total*	[native] *% of total*	total count	trusted *% of total*	(no overlap) *% of total*	[native] *% of total*
en		–			564	85.1%	(84.6%)	[84.6%]	664	83.6%	(42.8%)	[28.3%]	442	–	(82.3%)	[17.6%]
ja	370	83.8%	(83.8%)	[63.5%]	386	89.1%	(87.6%)	[63.0%]	472	94.9%	(10.8%)	[0.0%]	447	–	(44.3%)	[0.7%]

Evaluation Time

	Data Quality Control Mechanism											
	LOC+GOLD^{bw}+PAY			LOC+GOLD^{bw}			GOLD^{b}			NONE		
TRG	evaluation period	(grading time)	[avg. time per assignment]	evaluation period	(grading time)	[avg. time per assignment]	evaluation period	(grading time)	[avg. time per assignment]	evaluation period	(grading time)	[avg. time per assignment]
en		–		4.8 days	(04:30:13)	[00:39]	0.9 days	(03:24:45)	[00:25]	0.4 days	(01:12:32)	[00:17]
ja	0.2 days	(02:04:03)	[00:25]	12.8 days	(02:22:28)	[00:27]	0.7 days	(01:39:58)	[00:14]	0.3 days	(01:29:50)	[00:10]

5 Evaluation Results

In order to investigate the effects of data quality control mechanisms, the analysis of the evaluation results is conducted experiment-wise. i.e., we do not differentiate between single workers, but treat all collected judgments of the respective experiment as a "single" grader result. This enables a comparison of non-expert vs. expert grading results and of the impact of each control setting on the quality of collected judgments.

5.1 Worker Characteristics

Table 2 summarizes the characteristics of MTurk workers taking part in experiments, the amount of collected judgments and the evaluation time needed to carry out the MT evaluation. For each control setting, we list the amount of workers (*total*) and the percentage of (a) trusted workers (*trusted*), (b) trusted and non-overlapping[7] workers (*no overlap*), and (c) trusted and non-overlapping workers with origin in a country where the target language is the official language (*native*). Concerning the evaluation time, we measured the evaluation period, i.e., the number of days needed to collect the data, the grading time, i.e., the hours spent on actually grading the translations, and the average grading time per assignment.

The demographics of the MTurk workers, i.e. their country of origin, show that the judgments mainly originated from non-native workers with an average grading time of 10 (17) seconds per assignment for Japanese MT evaluation experiments without any control mechanism in place. The *GOLD^{b}* settings resulted in very high trust levels (65∼100%), but achieved worse figures with respect to non-overlap and native worker contributions. Concerning the evaluation time, more effort was spent on the task increasing the evaluation period by a factor of 3 and the overall grading time by a factor of 2.

For the *LOC+GOLD^{bw}* experiments, we limited the worker origin to the offical language countries and included the US for Japanese due to the expected lack of native speakers. In addition, we annotated both the best and worst translation of the control units. The results of the *LOC+GOLD^{bw}* control setting showed that the amount of judgments collected from native speakers increased by 63% and 56% for Japanese and English, respectively. However, the average time of the evaluation period increased by a factor of 18 (ja) and 5 (en) compared to the *GOLD^{b}* settings. For Japanese, it took almost 2 weeks to collect the data which may not be acceptable if time is a crucial factor.

[7]CF assigns unique IDs to each worker that can be used to trace which worker carried out HITs for which job. We exploit this feature to check whether a single worker took part in evaluation experiments for more than one target language.

Table 3: Ranking Results

	Data Quality Control Mechanism															
	LOC+GOLD^{bw}+PAY (MT Engine)				LOC+GOLD^{bw} (MT Engine)				GOLD^{b} (MT Engine)				NONE (MT Engine)			
TRG	160k	80k	20k	10k	160k	80k	20k	10k	160k	80k	20k	10k	160k	80k	20k	10k
en	–				0.4766	0.3481	0.2343	0.1138	0.2853	0.2620	0.1673	0.0750	0.1605	0.1714	0.1020	0.0680
ja	0.4864	0.3716	0.1786	0.0734	0.4811	0.3695	0.1461	0.0755	0.2355	0.1639	0.1281	0.0675	0.0724	0.0678	0.0470	0.0165

In order to measure how evaluation time relates to the evaluation costs, we repeated the Japanese MT evaluation experiment increasing the payment by a factor of 4. As a result, the *LOC+GOLD^{bw}+PAY* evaluation data could be collected within 5 hours indicating that *the more money paid, the sooner evaluation results can be expected to be available*.

5.2 Ranking Results

The *Ranking* scores were obtained as the average number of times that a system was judged better than any other system. The results summarized in Table 3 largely differ for the investigated data quality settings. The system rankings for workers in the uncontrolled tasks differ from the expert rankings for both target languages. However, the usage of the on-the-fly control mechanism resulted in the collection of more reliable judgments, ranking all four MT systems correctly. Interestingly, the ranking scores obtained for the better controlled evaluation experiments are much higher, indicating the collected evaluation data is of good quality.

The quality of the judgment is confirmed by the ranking agreement scores listed in Table 4. The self-consistency of uncontrolled data is extremely high indicating the usage of certain grading patterns resulting in unreliable judgments. *Substantial* agreement was mainly achieved by *LOC+GOLD^{bw}* workers. Comparing the worker vs. the expert judgments, only *slight* agreement was obtained for the less controlled settings, but the proposed data quality control mechanism achieved levels of up to *substantial* agreement.

For the *LOC+GOLD^{bw}+PAY* experiment, however, a lower percentage of trusted judgments was collected and only moderate agreement with expert judgments was achieved. This indicates that *increasing the pay does not necessarily increase the reliability of the evaluation data*.

Table 4: Ranking Agreement

Self-Consistency

$\mathcal{K}$intra	Data Quality Control Mechanism			
TRG	LOC+GOLD^{bw} +PAY	LOC+GOLD^{bw}	GOLD^{b}	NONE
en	–	0.54	0.27	0.78
ja	0.83	0.75	0.72	0.98

Worker vs. Expert Agreement

$\mathcal{K}$inter	Data Quality Control Mechanism			
TRG	LOC+GOLD^{bw} +PAY	LOC+GOLD^{bw}	GOLD^{b}	NONE
en	–	0.62	0.30	0.43
ja	0.58	0.66	0.22	0.23

6 Conclusions

In this paper we have investigated the use of various data quality control mechanisms of online work marketplaces to collect high-quality MT evaluation data for translations into English and Japanese. The analysis of the worker characteristics revealed that *locale qualification* control settings enable the collection of less tainted judgments and that bad workers can be identified by short HIT grading times and low trust levels measured on-the-fly during the evaluation task.

The improved setting of control units to verify not only the best but also the worst translation helped to identify untrusted workers using fixed gradings schemes. Finally, the combination of multiple control mechanism proved to be essential for collecting high-quality data in a reasonable period of time.

As future work, we are planning to investigate the applicability of the proposed crowd-based MT evaluation method to other non-English target languages and more complex translation tasks ranking more MT systems as well as covering other translation domains.

References

Bentivogli, L., M. Federico, G. Moretti, and M. Paul. 2011. Getting Expert Quality from the Crowd for MT Evaluation. In *Proceedings of the MT Summmit XIII*, pages 521–528.

Callison-Burch, C., C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proc. of the Second Workshop on SMT*, pages 136–158.

Callison-Burch, C. 2009. Fast, Cheap, and Creative: Evaluating MT Quality Using Amazon's Mechanical Turk. In *Proc. of the EMNLP*, pages 286–295.

Ipeirotis, P. 2010. New demographics of Mechanical Turk. http://hdl.handle.net/2451/29585.

Kikui, G., S. Yamamoto, T. Takezawa, and E. Sumita. 2006. Comparative study on corpora for speech translation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1674–1682.

Landis, J. and G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 (1):159–174.

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a Method for Automatic Evaluation of MT. In *Proc. of the 40th ACL*, pages 311–318.