Crowd-based MT Evaluation for non-English Target Languages

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

Luisa Bentivogli and **Marcello Federico**
FBK-irst
Via Sommarive, 18
38123 Povo-Trento, Italy
{bentivo,federico}@fbk.eu

Abstract

This paper investigates the feasibility of using crowd-sourcing services for the human assessment of machine translation quality of translations into *non-English* target languages. Non-expert graders are hired through the CrowdFlower interface to Amazon's Mechanical Turk in order to carry out a ranking-based MT evaluation of utterances taken from the travel conversation domain for 10 Indo-European and Asian languages. The collected human assessments are analyzed for their worker characteristics, evaluation costs, and quality of the evaluations in terms of the agreement between non-expert graders and expert/oracle judgments. Moreover, data quality control mechanisms including "locale qualification" "qualificatio testing", and "on-the-fl verification are investigated in order to increase the reliability of the crowd-based evaluation results.

1 Introduction

This paper focuses on the evaluation of machine translation (MT) quality for target languages other than English. Although human evaluation of MT output provides the most direct and reliable assessment, it is time consuming, costly, and subjective. Various automatic evaluation measures were proposed to make the evaluation of MT outputs cheaper and faster (Przybocki et al., 2008), but automatic metrics have not yet proved able to consistently predict the usefulness of MT technologies. To counter the high costs in human assessment of MT outputs, the usage of crowdsourcing services such as Amazon's Mechanical Turk[1] (MTurk) and CrowdFlower[2] (CF) were proposed recently (Callison-Burch, 2009; Callison-Burch et al., 2010; Denkowski and Lavie, 2010).

[1]http://www.mturk.com
[2]http://crowdfl wer.com

The feasibility of crowd-based MT evaluations was investigated for shared tasks such as the WMT (Callison-Burch, 2009) and the IWSLT (Federico et al., 2011) evaluation campaigns. Their results showed that agreement rates for non-experts were comparable to those for experts, and that the crowd-based rankings correlated very strongly with the expert-based rankings. Most of the crowd-based evaluation experiments focused on English as the target language, with the exception of (Callison-Burch et al., 2010) evaluating Czech, French, German, and Spanish translation outputs and (Federico et al., 2011) evaluating translations into French.

This paper investigates the feasibility of using crowdsourcing services for the human assessment of translation quality of translation tasks where the target language is *not* English, with a focus on non-European languages. In order to identify non-English target languages for which we can expect to fin qualifie workers, we referred to existing surveys that analyze the demographics of MTurk workers (see Section 2). In total, we selected 7 non-European languages consisting of Arabic (ar), Chinese (zh), Hindi (hi), Japanese (ja), Korean (ko), Russian (ru), and Tagalog (tl), as well as 3 European languages covering English (en), French (fr), and Spanish (es) as the target languages for our translation experiments.

The MT evaluation was carried out using utterances taken from the domain of travel conversations. A description of the utilized language resources and the MT engines are summarized in Section 3. 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).

For the 10 investigated language pairs, non-expert graders were hired through the CF interface to MTurk in order to carry out the ranking-based MT evaluation as described in Section 4. In addition, expert graders were employed for four of

the target languages (en, ja, ko, zh) to carry out exactly the same evaluation task as the non-expert workers. For all target languages without expert graders, we used an oracle ranking metric based on the "Training Size Preference" assumption, i.e., *the larger the training size, the better the translation quality can be expected to be*, to evaluate the quality of the worker judgments.

Besides a thorough analysis of the obtained non-expert grading results, we also investigated different data quality control mechanisms in order to increase the reliability of crowd-based evaluation results (see Section 5). The experiments carried out in this paper revealed that the quality of the crowd-based MT evaluation is closely related to the demographics of the online work marketplace. Although high-quality evaluation results could be collected for the majority of the investigated non-English languages, the need for multi-layered data quality control mechanisms causes an increase in evaluation time. The finding of this paper confi m that crowdsourcing is an effective way of reducing the costs of MT evaluation without sacrific ing quality even for non-English target languages given that control mechanisms carefully tailored to the evaluation task at hand are in place.

2 Mechanical Turk Demographics

Past surveys on the demographics of MTurk users indicated that most of the workers come from the US. (Ipeirotis, 2010) conducted a recent survey on the demographics of MTurk users which showed a shift in the "country of origin" of workers, i.e., a decrease in US workers to 47% and an increase of Indian workers to 34%, with the remaining 19% of workers coming from 66 different countries[3]. Based on the country information from MTurk workers taking part in the survey, we analyzed which languages are used by these workers.

The language distribution shows that the majority of workers speak English, followed by Hindi, Romanian, Tagalog, and Spanish. At least 5 workers were native speakers of Dutch, Arabic, Italian, German, and Chinese. However, taking into account officia languages spoken in the respective countries, we can expect larger contributions of workers speaking Spanish, French, and Arabic.

3 MT Evaluation Task

The crowd-based MT evaluation is carried out using the translation results of phrase-based statistical machine translation (SMT) systems that are trained on parallel corpora. The translation quality of SMT engines heavily depends on the amount of bilingual language resources available to train the statistical models. We exploited this characteristic of data-driven MT approaches to defin an "oracle" ranking metric (ORACLE) according to the "Training Size Preference" assumption, in which an MT output of a system A wins (or ties in) a comparison with the MT output of a system B, where the training corpus of system B is a subset of the one of system A.

The language resources used to build MT engines are described in Section 3.1. We selected 10 Indo-European and Asian languages based on the following criteria:

- "*Worker Availability*" covering languages with 'many' (en, hi), 'several' (es, tl), 'few' (ar, fr, ja, ru, zh), 'almost none' (ko) MTurk workers available.
- "*Usage for MT Research*" covering 'frequently' (ar, fr, zh), 'often' (es, ru), 'sporadically' (ja, ko) used languages as well as under-resourced languages (tl, hi).
- "*Availability of Language Resources*" used for the training and evaluation of MT engines.

The training corpus consisting of 160k relatively short sentences was split into three subsets of 80k, 20k, and 10k sentence pairs, respectively. Each subset was used to train an MT engine whose translation quality significantl differed from the others, with the MT engine trained on the full corpus achieving the best translation quality.

This translation experiment setup renders the manual evaluation relatively reliable due to (1) a relatively easy translation task and (2) large differences in translation performance between the utilized MT engines. Moreover, the ORACLE metric can be exploited to judge the quality of crowd-based evaluation results for all languages where expert graders were not available.

3.1 Language Resources

The crowd-based MT evaluation experiments are carried out using the multilingual *Basic Travel Expressions Corpus* (BTEC), which is a collection of sentences that bilingual travel experts consider useful for people going to or coming from another country (Kikui et al., 2006). The sentence-aligned corpus consists of 160k sentences and covers all 10 languages investigated in this paper.

The parallel text corpus was randomly split into three subsets: for evaluating translation quality (*eval*, 300 sentences), for tuning the SMT model weights (*dev*, 1000 sentences) and for training the

[3]Details on the survey can be found at http://hdl.handle.net/2451/29585

statistical models (*train*, 160k sentences). Furthermore, three subsets of varying sizes (80k, 20k, and 10k sentences) were randomly extracted from the training corpus and used to train four SMT engines on the respective training data sets for each of the investigated language pairs.

3.2 Translation Engines

The translation results evaluated in this paper were obtained using fairly typical phrase-based SMT engines built within the framework of a feature-based exponential model. For the training of the SMT models, standard word alignment (Och, 2003) and language modeling (Stolcke, 2002) tools were used. Minimum error rate training (MERT) was used to tune the decoder's parameters and was performed on the *dev* set using the technique proposed in (Och, 2003). For the translation, an in-house multi-stack phrase-based decoder was used.

In order to maximize the gains[4] from an increased training data size and therefore allow for reliable ORACLE judgments, we selected English as the source language for the translations into Arabic, Japanese, Korean, and Russian. For all other translation experiments, Japanese source sentences were used as the input for the SMT decoder.

3.3 Automatic Evaluation

For the automatic evaluation of translation quality, we applied the BLEU metric (Papineni et al., 2002). Scores range between 0 (worst) and 1 (best).

The results of the translation engines described in Section 3.2 are summarized in Table 1, where the BLEU scores are given as percent figu es (%BLEU). The obtained scores confi m the "Training Size Preference" assumption (160k>80k>20k>10k) of the ORACLE metric. Concerning the target languages, the highest BLEU scores were achieved for Korean and Japanese, followed by English, Chinese, Spanish and French. Arabic and Hindi seem to be the most difficul target languages for the given translation and evaluation tasks obtaining the lowest automatic evaluation scores for each of the investigated tasks.

3.4 Subjective Evaluation

Human assessments of translation quality were carried out using the *Ranking* metrics where human 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).

[4]For relatively simple translation tasks, the amount of training data affects the translation quality of closely related languages far less than for more distinct languages.

Table 1: Translation Quality (%BLEU)

Language		MT Engine			
Source	Target	160k	80k	20k	10k
en	ar	12.90	12.45	10.89	9.97
	ja	28.58	25.38	21.00	19.41
	ko	29.53	26.42	21.43	18.66
	ru	16.15	15.84	13.90	12.36
ja	en	24.47	19.95	15.35	12.57
	es	19.52	17.43	13.30	11.73
	fr	19.35	18.84	14.67	14.43
	hi	14.17	12.57	9.97	8.24
	tl	18.93	17.81	15.78	13.58
	zh	21.22	17.08	13.03	12.64

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 was changed randomly for each ranking set to avoid bias. 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*).

4 Crowd-based MT Evaluation

To counter the high costs in human assessment of MT outputs, crowdsourcing services such as MTurk and CF have attracted a lot of attention both from industry and academia as a means for collecting data for human language technologies 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 their quality. MTurk and CF provide requesters with quality control mechanisms including the "locale qualification option to restrict workers by country. Preliminary qualification for workers can be set by requiring workers to complete a qualifica tion test using training ranking sets. Only workers passing the test are allowed to accept a HIT for the evaluation task at hand. Moreover, CF provides a mechanism to verify the workers' reliability on-the-fl . The HIT design interface provided by CF allows including so called "gold units", i.e. items

with known labels, along with the other units composing the requested HIT. Gold units are randomly mixed with the other units by CF when it creates the worker assignments. These control units[5] allows 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 filte ed out from the fina data set. CF uses the workers' history to apply confidenc scores (the "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%. As a further control, CF pauses a job (the "auto-takedown" feature), if workers are failing too many gold units.

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 officia language countries; the most important control mechanism to prevent workers from tainting the evaluation results.
- *qualification testing* (PRI): training phase assessment of worker's eligibility prior to the evaluation task.
- *on-the-fly verification* (GOLD): identificatio of trusted workers using control units with a known answer.

4.2 Control Units

Control units have to be unambiguous, not too trivial, and also not too difficult For the translation task at hand, we selected the original corpus sentence as the main reference translation. From paraphrased reference translations[6], we selected a single reference as the *gold translation* to be included in the control units. A paraphrased reference to be selected as a gold translation should have the following characteristics: (1) it should be similar to the main reference and (2) its translation quality should be better than the best MT output for all translation hypotheses of the same input. If native speakers are available, the gold translation quality should be checked manually. However, for most of the investigated target languages, native speakers were not available. Thus, we automatically selected a gold translation based on the edit distance of each paraphrased reference to (a) the main reference and (b) the ORACLE-best (=160k) MT output for all sentence IDs of the *eval* set. We selected the most appropriate paraphrased reference according to its minimal distance to the main reference and its maximal distance to the MT output. 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 distinguished two GOLD annotation schemes:

- "best-only" (GOLD^{b}): check only the best translation, i.e., force rank '1' assignment for the gold translation.
- "best+worse" (GOLD^{bw}): check the best and the worst translation, i.e., allow rank '1' or '2' for the gold and rank '3' or '4' for the ORACLE-worst (10k) translation.

4.3 Evaluation Interface

CF provides two interfaces: (1) an *external* one for MTurk workers and (2) 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 from in-house expert graders using exactly the same HITs and the same online interface as the MTurk workers.

4.4 Experiment Setup

For each target language (TRG), we repeated the same MT evaluation experiment using the following data quality control settings[7]:

1. *NONE*: no quality control (all TRGs)
2. *GOLD*: on-the-fl only (all TRGs)
3. *LOC+GOLD*: locale+on-the-fl (all TRGs)
4. *LOC+GOLD+PRI*: locale+testing+on-the-fl (hi, ko)

All experiments using the same control setting were carried out simultaneously, i.e., a single worker might take part in more than one evaluation experiment. A HIT consisted of 3 ranking sets per page and is paid 6 cents for all experiments. In total, the evaluation costs[8] for all the experiments added up to $390 for 30 experiments, resulting in an average of $13 for the crowd-based evaluation of 4 MT outputs for 300 input sentences.

5 Evaluation Results

In order to investigate the effects of the 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 the collected judgments of the respective experiment as a "single" grader result. This enables a

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

[6]Up to 15 paraphrased reference translations are available for the data sets described in Section 3.1.

[7]India was excluded by default for all experiments besides the ones having Hindi as the target language.

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

comparison of non-expert vs. expert/oracle grading results and the impact of each control setting on the quality of the collected judgments. The details of the experiment results for each target language are listed in Appendix A.

5.1 Worker Characteristics

Table A.1. summarizes the amount of participating workers. For each control setting, we list the amount of workers (*total*) and the percentage of workers coming from a country where the language is the officia language (*native*). The worker demographics are summarized in Table A.2.

Without any control mechanism in place, the judgments mainly originated from non-native workers. 53% of the workers submitted HITs for at least two tasks, with the largest overlap being f ve tasks. Although some workers might be able to speak and evaluate more than two languages, the results indicate that *the larger the overlap, the less reliable the judgments are expected to be*.

The on-the-fl verificatio based on gold translations only ($GOLD^b$) resulted in a high percentage of judgments obtained from trusted workers (65∼100%) for the majority of tasks, but achieved worse figu es with respect to native worker contributions. These finding indicate that *single gold translations are not sufficient to identify workers assigning grades based on fixed patterns*.

As a counter-measure, we limited the worker origin to the officia language countries and the US, and annotated both the best and worst translation of the control units. As a results, 47% of the $LOC{+}GOLD^{bw}$ gradings were collected from native speakers. These results show that the locale and on-the-fl control enable the collection of less tainted judgments and the identificatio of untrusted workers, respectively. Table A.3. summarizes the amount of judgments collected for each task. The total count depends on the number of non-trusted workers accepting HITs for the respective language.

Although high-quality control units positively affect the quality of the evaluations as shown in Section 5.2, the average time needed to collect the data increased by a factor of 8. The evaluation period, i.e., the number of days needed to collect all the data, the grading time, i.e., the hours spent on actually grading the translations, and the average grading time per assignment are summarized in Table A.4. The grading time for each task ranged from 2.5h to 6.5h for the $LOC{+}GOLD^{bw}$ experiments. However, the evaluation period largely depends on the language, ranging from 2 days (hi, tl, es) to over 2 weeks (ru, zh, ko). The analysis of the average time needed to judge a single HIT indicates that *the shorter the evaluation time, the less reliable the judgments are expected to be*.

The most problematic languages are Korean and Hindi. For Korean, the evaluation experiments lasted 3 months due to the lack of trusted workers. Moreover, the Hindi $LOC{+}GOLD^{bw}$ task could not be finishe because the large amount of untrusted workers triggered CF's *auto-takedown* feature. In order to prevent an auto-takedown for jobs where low trust levels of workers are to be expected, a training phase assessing the worker's eligibility prior to the evaluation task needs to be included. Only workers passing the qualificatio test were allowed to accept HITs for the respective task. The Korean and Hindi results given in Appendix A were therefore obtained using the $LOC{+}GOLD^{bw}{+}PRI$ data quality control setting.

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 A.5. differ largely for the investigated data quality settings. System ranking scores resulting in an MT system ordering other than the expert rankings are marked in boldface. For most of the uncontrolled tasks, worker rankings are different from expert rankings. The $GOLD^b$ setting tasks achieved a higher correlation with expert rankings, but still differ for 3 out of the 10 languages. The $LOC{+}GOLD^{bw}$ tasks ranked all the MT systems identically to the experts. Interestingly, the ranking scores obtained for the better controlled evaluation experiments are much higher, indicating the collected evaluation data is of good quality.

5.3 Grading Consistency

The most informative indicator of the quality of a dataset is given by the agreement rate, or grading consistency, both between different judges and the same judge. To this purpose, the agreement between non-expert graders of experiments using different data quality control mechanisms was calculated for the MTurk data and compared to the results obtained by expert/oracle judgments. Agreement rates are calculated using the *Fleiss' kappa coefficient* κ (Fleiss, 1971):

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

where Pr(a) is the observed agreement among graders, and Pr(e) is the hypothetical probability of

chance agreement. In our task, Pr(a) is given by the proportion of times that two judges assessing the same pair of systems on the same source sentence agree that A>B, A=B, or A<B. Grader agreement scores can be interpreted as follows: "*none*" κ <0, "*slight*" κ ≤0.2, "*fair*" κ ≤0.4, "*moderate*" κ ≤0.6, "*substantial*" κ ≤0.8, and "*almost perfect*" κ ≤1.0 (Landis and Koch, 1977).

The quality of the judgment is confi med by the ranking agreement scores listed in Table A.6. Comparing the worker vs. the expert judgments, only *slight* agreement was obtained for the less controlled settings, but the proposed data quality control mechanisms achieved levels of up to *substantial* agreement. The comparison of agreement scores for oracle and expert judgments indicates that at least *fair* agreement is to be expected for languages where expert graders are not available.

6 Conclusions

In this paper, we investigated the use of the data quality control mechanisms of online work marketplaces for the collection of high-quality MT evaluation data for non-English target languages. 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 identifie by short HIT grading times, large overlaps of evaluation tasks run simultaneously, and low trust levels measured either prior to or during the evaluation task.

Due to the lack of expert graders for 6 out of 10 languages, the creation of control units was carried out automatically, where the proposed similarity-based gold translation selection method proved to be a practical alternative to manual selection by native speakers. The improved setting of control units to verify not only the best but also the worst translation helped to identify untrusted workers using fi ed gradings schemes. Finally, the combination of multiple control mechanism proved to be essential for collecting high-quality data for all the investigated non-English languages.

Based on the obtained findings we recommend carrying out crowd-based MT evaluations by (1) limiting the access to workers in countries where the target language is the officia language, although for languages lacking workers, the US might be included if evaluation time is a crucial factor and (2) definin control units so that expected rankings for the best and the worst systems are preserved and grading variations of non-expert graders are taking into account.

As future work, we are planning to investigate the effectiveness of other control mechanisms such as *payment* and the applicability of the proposed crowd-based MT evaluation method to more complex translation tasks, ranking more MT systems, as well as covering other domains such as the translation of public speeches.

References

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., P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan. 2010. Findings of the 2010 Joint Workshop on SMT and Metrics for MT. In *Proc. of the Joint Fifth Workshop on SMT and MetricsMATR*, pages 17–53.

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

Denkowski, M. and A. Lavie. 2010. Exploring Normalization Techniques for Human Judgments of Machine Translation Adequacy Collected Using Amazon Mechanical Turk. In *Proc. of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 57–61.

Federico, M., L. Bentivogli, M. Paul, and S. Stücker. 2011. Overview of the IWSLT 2011 Evaluation Campaign. In *Proc. of IWSLT*, pages 11–27.

Fleiss, J. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76 (5):378–382.

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.

Och, F.J. 2003. Minimum Error Rate Training in SMT. In *Proc. of the 41st ACL*, pages 160–167.

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.

Przybocki, M., K. Peterson, and S. Bronsart. 2008. Metrics for MAchine TRanslation Challenge. http://nist.gov/speech/tests/metricsmatr/2008/results.

Stolcke, A. 2002. SRILM: an extensible language modeling toolkit. In *Proc. of the ICSLP*.

Appendix A. Crowd-based MT Evaluation

A.1. Amount of Workers

The total number of participating workers, as well as the number and the percentage of trusted/native workers for each evaluation task.

TRG	Data Quality Control Mechanism: LOC+GOLD^{bw} total count	LOC+GOLD^{bw} trusted *(% of total)*	LOC+GOLD^{bw} native *[% of total]*	GOLD^{b} total count	GOLD^{b} trusted *(% of total)*	GOLD^{b} [native] *[% of total]*	NONE total count	NONE trusted *(% of total)*	NONE native *[% of total]*
en	23	18 (78.3%)	13 [56.5%]	38	30 (78.9%)	10 [26.3%]	8	–	4 [50.0%]
ar	41	26 (63.4%)	23 [56.0%]	29	19 (65.5%)	6 [20.6%]	14	–	**0 [0.0%]**
es	19	19 (100.0%)	15 [78.9%]	12	11 (91.6%)	2 [16.6%]	8	–	**0 [0.0%]**
fr	10	9 (90.0%)	4 [40.0%]	10	9 (90.0%)	**0 [0.0%]**	14	–	2 [14.2%]
hi	31*	28* (90.3%)	27* [87.0%]	85	37 (43.5%)	34 [40.0%]	47	–	33 [70.2%]
ja	14	11 (78.5%)	3 [21.4%]	15	13 (86.6%)	**0 [0.0%]**	10	–	1 [10.0%]
ko	45*	43* (95.5%)	2* [4.4%]	24	17 (70.8%)	**0 [0.0%]**	5	–	**0 [0.0%]**
ru	30	20 (66.6%)	4 [13.3%]	7	7 (100.0%)	**0 [0.0%]**	14	–	**0 [0.0%]**
tl	10	9 (90.0%)	5 [50.0%]	6	6 (100.0%)	**0 [0.0%]**	2	–	1 [50.0%]
zh	18	11 (61.1%)	3 [16.6%]	16	12 (75.0%)	**0 [0.0%]**	7	–	**0 [0.0%]**

* marked results are obtained using the *LOC+GOLD^{bw}+PRI* data quality control setting.

A.2. Country of Origin

The total number of countries and workers per country participating in each evaluation task.

TRG	Data Quality Control Mechanism: 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
ar	11 countries JOR:12, EGY:8, USA:7, TUN:3, LBN:3, SAU:2, MAR:2, DZA:1, KWT:1, ARE:1, OMN:1	15 countries MKD:6, TUN:3, JOR:3, EGY:2, USA:2, BGD:2, ARE:2, GBR:2, DZA:1, CHN:1, ESP:1, MDV:1, ROU:1, OMN:1, SAU:1	10 countries MKD:3, EGY:2, PAK:2, CHN:1, DZA:1, GBR:1, LBN:1, TUN:1, ARE:1, USA:1
es	8 countries ESP:5, MEX:4, USA:4, COL:2, ARG:1, GTM:1, URY:1, VEN:1	5 countries MKD:7, ESP:2, USA:1, BGD:1, ROU:1	7 countries USA:2, BHS:1, ESP:1, PRT:1, MKD:1, PAK:1, ROU:1
fr	4 countries USA:5, FRA:3, CAN:1, CMR:1	5 countries MKD:6, USA:1, CMR:1, NLD:1, ROU:1	8 countries MKD:3, PAK:3, FRA:2, ROU:2, CAN:1, CMR:1, NLD:1, USA:1
hi	2 countries IND:30*, USA:1*	4 countries IND:80, PAK:3, USA:1, ROU:1	8 countries IND:33, MKD:6, CHN:2, PAK:2, SGP:1, ARE:1, ROU:1, USA:1
ja	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
ko	2 countries USA:41, KOR:2	10 countries MKD:9, ROU:3, PHL:3, USA:2, CHN:2, POL:1, BGD:1, MDV:1, PAK:1, ESP:1	3 countries CHN:2, USA:2, MKD:1
ru	2 countries USA:25, RUS:5	5 countries PAK:2, ROU:2, GBR:1, SRB:1, MKD:1	7 countries MKD:8, MDA:1, POL:1, SRB:1, UKR:1, CHN:1, PAK:1
tl	2 countries PHL:7, USA:3	3 countries MKD:3, ROU:2, PAK:1	1 country PHL:2
zh	4 countries USA:12, CHN:3, SGP:2, HKG:1	6 countries MKD:9, USA:3, ROU:1, NLD:1, CHN:1, BGD:1	4 countries USA:3, CHN:2, SGP:1, MKD:1

* marked results are obtained using the *LOC+GOLD^{bw}+PRI* data quality control setting.

A.3. Judgments

The total number of rankings sets judged by all/trusted/native workers for each evaluation task.

TRG	Data Quality Control Mechanism: LOC+GOLD^{bw} total count	LOC+GOLD^{bw} trusted *(% of total)*	LOC+GOLD^{bw} native *[% of total]*	GOLD^{b} total count	GOLD^{b} trusted *(% of total)*	GOLD^{b} native *[% of total]*	NONE total count	NONE trusted *(% of total)*	NONE native *[% of total]*
en	564	495 (87.8%)	168 [29.8%]	664	568 (85.5%)	128 [19.3%]	442	–	78 [17.6%]
ar	693	543 (78.4%)	432 [62.3%]	559	463 (82.8%)	117 [20.9%]	465	–	**0 [0.0%]**
es	581	581 (100.0%)	542 [93.3%]	428	416 (97.2%)	86 [20.1%]	421	–	**0 [0.0%]**
fr	463	409 (88.3%)	178 [38.4%]	416	404 (97.1%)	**0 [0.0%]**	495	–	18 [3.6%]
hi	580*	505* (87.1%)	496* [85.5%]	1013	531 (52.4%)	477 [47.1%]	723	–	314 [43.5%]
ja	386	356 (92.2%)	60 [15.5%]	472	448 (94.9%)	**0 [0.0%]**	447	–	**0 [0.0%]**
ko	642*	603* (93.9%)	66* [10.3%]	583	523 (89.7%)	**0 [0.0%]**	408	–	**0 [0.0%]**
ru	657	555 (84.5%)	96 [14.6%]	370	370 (100.0%)	**0 [0.0%]**	504	–	**0 [0.0%]**
tl	437	428 (97.9%)	91 [20.8%]	344	344 (100.0%)	**0 [0.0%]**	371	–	36 [9.7%]
zh	575	481 (83.6%)	354 [61.6%]	462	429 (92.9%)	**0 [0.0%]**	476	–	**0 [0.0%]**

* marked results are obtained using the *LOC+GOLD^{bw}+PRI* data quality control setting.

A.4. Evaluation Time

The evaluation period (given in days), the total grading time (given in hours, "(hh:mm:ss)"), and the average time per HIT (given in seconds, "[mm:ss]") of the trusted gradings obtained for each evaluation task.

	Data Quality Control Mechanism											
	EXPERT			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	6.9 days	(06:41:09)	[00:13]	4.8 days	(04:30:13)	[00:39]	0.9 days	(03:24:45)	[00:25]	0.4 days	(01:12:32)	[00:17]
ar		–		4.7 days	(06:29:32)	[00:47]	0.7 days	(02:48:34)	[00:24]	0.1 days	(00:45:50)	[00:07]
es		–		2.2 days	(06:06:55)	[00:47]	0.3 days	(01:49:34)	[00:16]	0.1 days	(00:48:22)	[00:07]
fr		–		3.9 days	(04:19:36)	[00:40]	0.2 days	(03:13:52)	[00:29]	0.2 days	(00:55:04)	[00:14]
hi		–		1.2 days*	(03:27:34)*	[00:35]*	0.2 days	(02:44:42)	[00:19]	0.1 days	(00:52:55)	[00:08]
ja	1.1 days	(05:48:35)	[01:07]	12.8 days	(02:22:28)	[00:27]	0.7 days	(01:39:58)	[00:14]	0.1 days	(01:07:17)	[00:10]
ko	7.1 days	(11:29:41)	[00:16]	88.9 days*	(04:45:05)*	[00:41]*	3.1 days	(01:10:46)	[00:10]	0.1 days	(01:07:52)	[00:11]
ru		–		17.0 days	(06:48:44)	[00:52]	0.1 days	(01:55:05)	[00:18]	0.2 days	(01:12:47)	[00:11]
tl		–		2.1 days	(03:03:16)	[00:26]	0.1 days	(00:43:59)	[00:07]	0.1 days	(01:07:17)	[00:10]
zh	1.1 days	(07:32:56)	[01:26]	23.7 days	(05:09:30)	[00:43]	2.1 days	(01:29:36)	[00:13]	0.1 days	(01:52:16)	[00:16]

* marked results are obtained using the *LOC+GOLD^{bw}+PRI* data quality control setting.

A.5. Ranking Results ($\%_{better}$)

The subjective evaluation of translation quality of 4 MT engines trained on different training data sizes (160k, 80k, 20k, 10k). The *Ranking* scores were obtained as the average number of times that a system was judged better than any other system.

	Data Quality Control Mechanism															
	EXPERT				LOC+GOLD^{bw}				GOLD^{b}				NONE			
TRG	160k	80k	20k	10k	160k	80k	20k	10k	160k	80k	20k	10k	160k	80k	20k	10k
en	0.5245	0.4755	0.3272	0.1453	0.4766	0.3481	0.2343	0.1138	0.2853	0.2620	0.1673	0.0750	**0.1605**	**0.1714**	0.1020	0.0680
ar		–			0.4319	0.3038	0.1943	0.1497	0.1816	0.1135	0.0837	0.0723	**0.0008**	**0.0009**	**0.0019**	**0.0081**
es		–			0.4899	0.4062	0.2342	0.1176	0.1983	0.1474	0.0758	0.0620	**0.0000**	**0.0000**	**0.0000**	**0.0000**
fr		–			0.4823	0.4020	0.1652	0.0908	0.1929	0.1631	**0.0879**	**0.1035**	0.0400	**0.0326**	**0.0370**	**0.0370**
hi		–			0.2837*	0.2068*	0.1094*	0.0889*	0.1872	0.1587	**0.0868**	**0.0947**	0.0201	**0.0111**	**0.0191**	0.0040
ja	0.5735	0.4803	0.2528	0.1027	0.4811	0.3695	0.1461	0.0755	0.2355	0.1639	0.1281	0.0675	0.0724	0.0678	0.0470	0.0165
ko	0.4690	0.3746	0.2625	0.1136	0.3809*	0.3185*	0.1740*	0.0919*	0.0862	0.0689	0.0532	0.0517	**0.0000**	**0.0000**	**0.0000**	**0.0000**
ru		–			0.3459	0.2957	0.1830	0.1078	0.2588	0.2390	0.1887	0.1613	0.0606	0.0552	0.0433	0.0400
tl		–			0.3914	0.2679	0.1428	0.1027	0.0679	0.0648	**0.0340**	**0.0340**	0.0022	**0.0011**	**0.0022**	**0.0044**
zh	0.5482	0.4313	0.3318	0.2133	0.6367	0.5128	0.4110	0.2811	0.1371	0.1331	0.1223	0.1035	0.0802	0.0552	0.0542	0.0427

* marked results are obtained using the *LOC+GOLD^{bw}+PRI* data quality control setting.

A.6. Ranking Agreement

Fleiss' kappa correlation coefficient comparing the obtained crowd-based evaluation results to the oracle and expert judgments for each translation task. The κ scores are interpreted in (Landis and Koch, 1977) as follows:

$\kappa < 0$: "*none*" $\kappa \leq 0.6$: "*moderate*"
$\kappa \leq 0.2$: "*slight*" $\kappa \leq 0.8$: "*substantial*"
$\kappa \leq 0.4$: "*fair*" $\kappa \leq 1.0$: "*almost perfect*"

Worker vs. Oracle/Expert Agreement

κ	**Data Quality Control Mechanism**					
	LOC+GOLD^{bw}		GOLD^{b}		NONE	
TRG	oracle	expert	oracle	expert	oracle	expert
en	0.45	0.62	0.19	0.30	0.39	0.43
ar	0.22	–	0.09	–	0.11	–
es	0.35	–	0.08	–	1.00	–
fr	0.26	–	0.04	–	0.53	–
hi	0.05*	–	0.00	–	-0.02	–
ja	0.38	0.66	0.10	0.22	0.01	0.23
ko	0.56	0.50	0.79	0.14	-0.01	0.17
ru	0.32	–	0.08	–	0.15	–
tl	0.21	–	0.04	–	-0.02	–
zh	0.62	0.56	0.07	0.09	0.17	0.20

* marked results are obtained using the *LOC+GOLD^{bw}+PRI* data quality control setting.