Crowd-Sourcing of Human Judgments of Machine Translation Fluency

Yvette Graham **Timothy Baldwin** **Alistair Moffat** **Justin Zobel**
Department of Computing and Information Systems, The University of Melbourne
{ygraham,tbaldwin,ammoffat,jzobel}@unimelb.edu.au

Abstract

Human evaluation of machine translation quality is a key element in the development of machine translation systems, as automatic metrics are validated through correlation with human judgment. However, achievement of consistent human judgments of machine translation is not easy, with decreasing levels of consistency reported in annual evaluation campaigns. In this paper we describe experiences gained during the collection of human judgments of the fluency of machine translation output using Amazon's Mechanical Turk service. We gathered a large collection of crowd-sourced human judgments for the machine translation systems that participated in the WMT 2012 shared translation task, collected across a range of eight different assessment configurations to gain insight into possible causes of – and remedies for – inconsistency in human judgments. Overall, approximately half of the workers carry out the human evaluation to a high standard, but effectiveness varies considerably across different target languages, with dramatically higher numbers of good quality judgments for Spanish and French, and the reverse observed for German.

1 Introduction

The ability to accurately measure the properties of an object of study, such as a computational system, is fundamental to progress in science. For measurements to be meaningful, they need to be comparable between systems, and to be an accurate proxy for the properties of the systems being studied.

For machine translation (MT), measurement has been a combination of human judgments and automated measurements. With the aim of removing system biases and creating robust comparisons, there has been extensive use of workshops and shared tasks such as the ongoing Workshops on Statistical Machine Translation (WMT) and the NIST Open Machine Translation (OpenMT) evaluations. The basis of system evaluation is generally human judgments, which have also been used to evaluate automatic metrics such as BLEU (Papineni et al., 2001), under the assumption that a metric that correlates strongly with human judgments is more valid than a metric with weak correlation. Human evaluation of MT thus forms the foundation of evaluation in empirical MT, regardless of whether a particular evaluation makes use of human judges or automatic metrics.

The current methodology used for the task of human evaluation in MT is problematic, however, as assessments carried out by expert judges are highly inconsistent. Even when a single expert judge is asked to assess the same pair of translations in two separate sittings, the second judgment is often at odds with the initial one (Bojar et al., 2013). Somewhat paradoxically, and despite the fact that experts are not consistent, when non-experts are employed to do judgments, there is a tendency to give preference to non-experts who demonstrate high agreement with experts.

We have used Amazon's Mechanical Turk service (AMT) to gather human judgments of machine translations. Here we describe the data we have collected, our experiences in gathering this data, and our refinements to the gathering process. In particular, we have carried out a large-scale human evaluation across a range of different assessment configurations. The following assessment dimensions were explored: response scale; question wording; whether to include a reference translation; and deletion of foreign language words from translations. To ensure that the results are not peculiar to a single language pair, we in-

Yvette Graham, Timothy Baldwin, Alistair Moffat and Justin Zobel. 2013. Crowd-Sourcing of Human Judgments of Machine Translation Fluency. In *Proceedings of Australasian Language Technology Association Workshop*, pages 16−24.

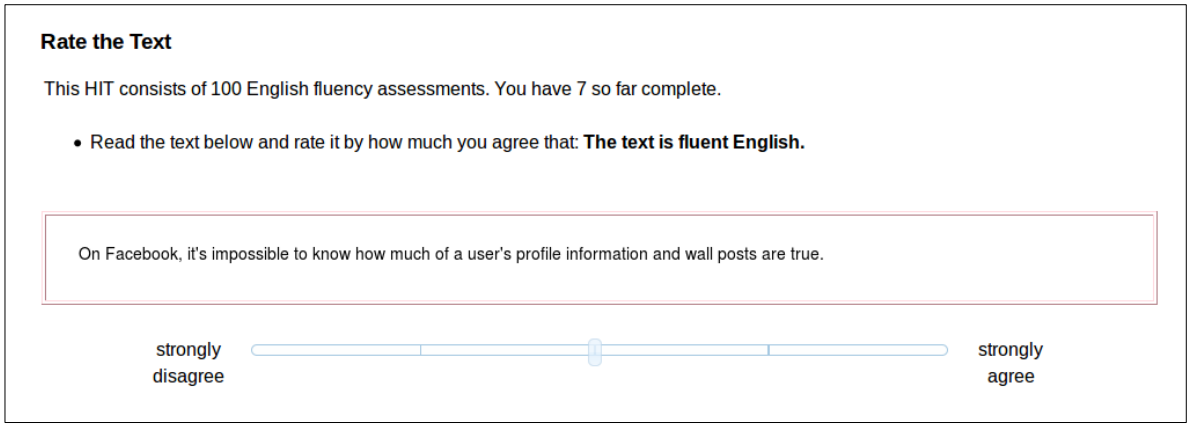

Figure 1: Screen shot for *base* configuration for fluency assessments, including 100 point visual analog scale (VAS), marked but not labeled at 25-50-75.

clude seven language pairs across all participating systems from WMT 2012 (Callison-Burch et al., 2012).

Previous work has shown the advantages of collecting judgments on a continuous rating scale for NLP evaluation (Belz and Kow, 2011) in general, as well as for MT evaluation specifically (Graham et al., 2013), as shown in Figure 1. This approach allows judge-intrinsic quality control to be introduced, so that non-experts can be used, as well as permitting standardization of scores and longitudinal evaluation. We adopt this approach and ask AMT workers to assess the fluency of translations on a continuous rating scale. Since we are primarily concerned with design of the assessment configuration so as to improve the consistency of human judgments, and not with ranking of systems, we limit our assessment to evaluating *fluency*. Graham et al. (2012) suggest *translation quality* should be measured as a hypothetical construct, where measurements that employ more items (dimensions of measurement) as opposed to fewer are considered more valid. Under this criterion, a two-item (fluency and adequacy) scale is more valid than a single-item translation quality measure, further motivating the inclusion of fluency as an assessment item for measurement of translation quality.

Overall, just under half of the Turkers carried out the human evaluation to a standard that met our quality control threshold. In addition, proportions of good quality workers vary considerably from one target language to the next, with dramatically higher proportions of good quality judgments for Spanish and French. The reverse occurs for translation into German, however, where less than one third of completed Human Intelligence Tasks (HITs) were carried out by workers that reached the quality control threshold.

2 Assessment Design

The data we have gathered explores four dimensions of MT quality assessment for fluency: question wording; labeling of the response scale; inclusion of reference translation; and presence of source language words in translations. We first establish a *base configuration* assessment set-up from which seven other configurations are created. Figure 1 shows a screen shot of the base assessment configuration.

For each variant configuration, a single dimension of the base configuration is changed, as shown in Figure 2. The same 100-point continuous response scale was used for all configurations, based on the findings of Graham et al. (2013). All configurations were then applied to seven language pairs. In all cases, instructions and questions were presented to the judges in the target language.

The first dimension of the assessment design we investigate is alternative possible anchor labels of the *visual analog scale* (VAS). The scale shown in Figure 1 is the base assessment configuration, and uses a 100-point *marked VAS* response scale, with tick marks at 25, 50, and 75. Two variants were also explored (the "east" dimension in Figure 2): an unmarked VAS, which omits the markings on the response scale (shown at the top of Figure 3);

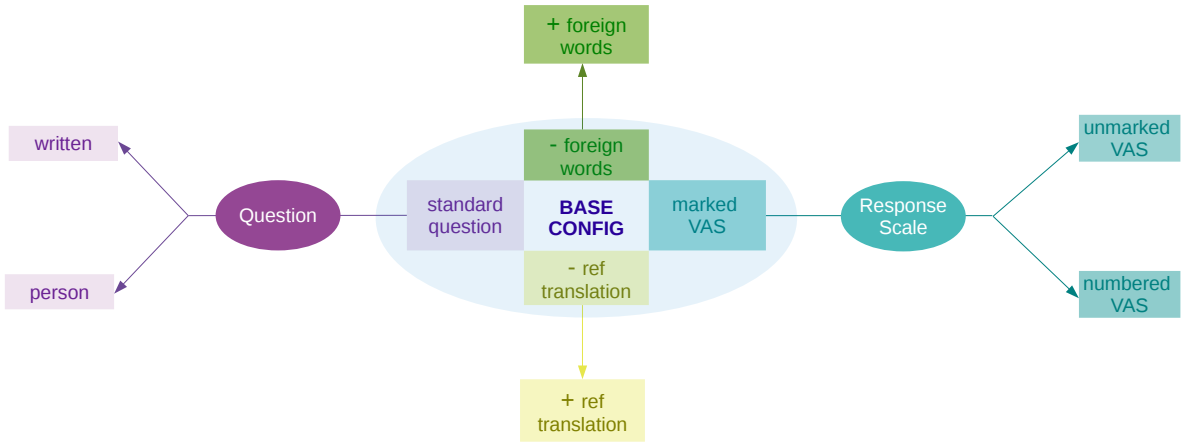

Figure 2: Trialed fluency assessment configurations: base configuration (center); additional assessment configurations diverge from the base on a single dimension.

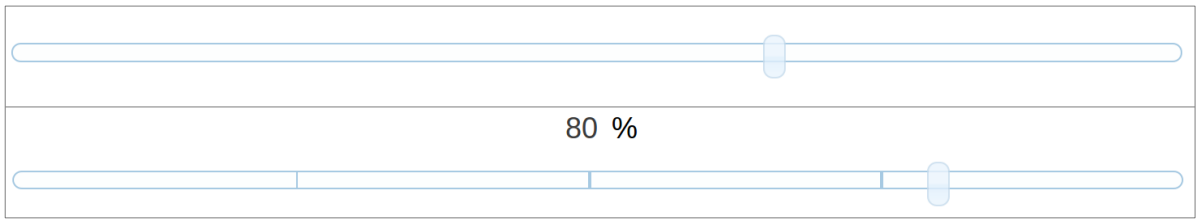

Figure 3: Variant VAS arrangements: unmarked and unlabeled, top; and marked and numbered VAS (100 point scale, marked at 25-50-75, displaying percentage corresponding to slider position), bottom.

and a numbered VAS that provides the judge with the numeric position at which the slider is sitting, a value that smoothly changes when the slider is moved (shown at the bottom of Figure 3).

When asking human judges to assess the fluency of translations, the particular way in which the question is asked is of obvious importance. The data we have collected includes trials of three alternative question wordings and response scale anchor labels (the "west" dimension in Figure 2); the three variants are shown in Table 1. First, the base configuration question (denoted *standard*) is a straightforward Likert declarative statement that directly uses the term "fluent English". But in everyday language usage, the term *fluent* is typically used to describe a person as opposed to expression. Hence, asking the judge whether the person who wrote the text is fluent might make the question more intuitive, and subsequently yield more consistent judgments – the *person* approach listed second in the table. Finally, we choose a wording that simply replaces "fluent" with a phrase more commonly used to refer to language, that is, whether the text is clearly written, denoted in Table 1 as the *written* approach.

The third dimension is whether or not to include a reference translation (the "south" variant in Figure 2). An assessment of fluency independent of adequacy and without a reference translation provides at least one part of an overall evaluation that will not be biased in favor of systems that happen to produce reference-like translations. However, in the past, fluency judgments have generally been carried out with a reference translation present (Callison-Burch et al., 2007). In this part of the evaluation the instructions described the task as assessing automatic translations as opposed to a simple rating of the fluency of the text, since without this context it would be difficult to explain what a reference in fact was. With each translation that was presented a note was displayed on screen to the users as follows: *An equivalent piece of fluent text is provided in gray for your reference.*

The final dimension explored (the "north" variant in Figure 2) is the effect of the presence of source language words in translations. Many of

Configuration	Question	Anchor labels	
		left	right
standard	Read the text below and rate it by how much you agree that: **The text is fluent English**.	strongly disagree	strongly agree
person	Based only on the text below, estimate the extent to which **the person who wrote the text is fluent in English**.	not at all fluent	highly fluent
written	**Is the text in the box below clearly written?**	not at all clear	very clear

Table 1: Alternative wordings for the instructions given to Turkers.

the translations in the data set contain foreign language words, due to MT systems whose response to words that are unknown is to leave them untranslated in the output. The presence of foreign words could be a cause of inconsistency for human judges, however. If, for example, a human judge happens to know the source language, their assessment of a translation containing a foreign word might be more favorable than that of a judge who has no knowledge of the source language. We therefore carried out an assessment with foreign words removed from translations (in the *base* configuration), and a contrasting assessment where foreign words were retained.

3 Data Set

The data set we have gathered consists of approximately 91,000 human judgments of the fluency of translations drawn from the WMT 2012 shared task published data (Callison-Burch et al., 2012). For each of the language pairs German-English, French-English, Spanish-English, Czech-English, English-German, English-French, and English-Spanish 560 system outputs were selected at random across participating systems. To each set of translations, an additional 240 translations were added as same-judge quality-control repeat items, 80 of which were exact repeats (*ask_again*) of a previously assessed translation, another 80 a *bad-reference* item, and the final 80 were reference translations, which should be judged highly by all judges. Thus for each language pair, a set of 800 translations was assessed across the eight different assessment configurations. (We also sought English-Czech judgments, but received a low response rate at AMT.)

Same Judge Repeat Items Control of same-judge repeat items on AMT with the conventional set-up is not straightforward, as a HIT usually consists of a single assessment (whether it be 5 translations or 1 translation per screen). To counter this, we use the unconventional HIT structure described by Graham et al. (2013) and constructed HITs of 100 judgments, so that we can fully control same-judge repeat items. We include a minimum number of 40 intervening judgments between repeat items, making it unlikely that a worker could boost their consistency by simply remembering a previous score.

Distinct Judge Repeat Items Control of distinct judge repeat items on AMT is straightforward, as the requester can specify for a set of HITs that they require a particular number of distinct workers. Since our focus is not on evaluating individual systems, but rather examining consistency of judgments, we specified that two distinct workers should carry out each HIT that we provided.

Worker Reliability We include in the data set for each AMT worker an estimate of their reliability based on score distributions for *bad-reference* pairs (explained below). The reliability estimate is a simplification of the method used in Graham et al. (2013) for quality-control. Instead of applying difference of means tests to *score differences* between that of the first and repeated item, we apply the same test to the mean of raw scores of *bad-reference* pairs.

No judge, when given the same translation to judge twice on a continuous scale (when separated by intervening judgment requests, the approach used in our experiments) can be expected to give precisely the same score for each judgment. A more flexible tool is thus required. We build such a tool by starting with two core assumptions:

A: When a consistent judge is presented with a set of repeat judgments, the mean score for the initial assessments will be neither significantly greater than nor significantly less than the mean score for repeat assessments.

Item A	Item B	MAE	κ_{intra}	κ_{inter}
47.1	47.2	13.3	0.68	0.37

Table 2: Agreement for *ask_again* repeat items for good workers.

Item A	Item B	MoD
26.0	47.3	21.3

Table 3: Agreement for *bad-reference* repeat items for good workers.

B: When a consistent judge is presented with a set of judgments for translations from two distinct systems, one of which is known to be better than the other, the mean score for the better system will be significantly higher than the mean score for the inferior system.

Assumption B is the basis of our reliability estimate, and allows us to distinguish between Turkers who are working carefully from those who are merely going through the motions. Deliberately degraded translations – referred to as *bad-reference* strings – are constructed from systems' translations and placed in to each HIT. Fluency-degraded translations were generated as follows: two words in the translation were randomly selected and randomly re-inserted elsewhere (but not as the initial or final word of the sentence). All translations, from all participating systems, were used to create *bad-reference* pairs, with a random subset used in HITs.

To compute the reliability estimate, *bad-reference* pair scores for a worker's HITS were extracted, a difference of means test undertaken, and the resulting p-value then used as a reliability estimate. A threshold (for example, $p < 0.05$) can then be applied to select the reliable workers. Careless judges have a high p-value, while judges who are both skilled and conscientious have a low p-value. This relationship can be validated by direct inspection of the judgments performed.

4 Judge Consistency

Table 2 shows consistency of human judges for judgments of translations repeated by the same and distinct judges. Mean scores for same judge *ask_again* repeat items show no significant difference. At the same time, mean scores for degraded *bad-reference* translations (Table 3) are significantly lower than for the corresponding system outputs. The Mean Absolute Error, computed as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \ , \qquad (1)$$

where f_i denotes the prediction (repeated score) and y_i the target (initial score) is 13.3 for *ask_again* items, while Mean of Differences, given by

$$\text{MoD} = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i) \ , \qquad (2)$$

for *bad-reference* repeat items is 21.3. Calculated Kappa coefficients for same- and distinct- judge repeat items are 0.68 and 0.37 respectively when the continuous scores are mapped to two categories: less than or equal to 50; and greater than 50. (Kappa values of 0.0–0.2 represent "slight" agreement; of 0.2–0.4 are "fair"; of 0.4–0.6 are "moderate"; of 0.6–0.8 are "substantial"; and of 0.8–1.0 represent agreements that are "almost perfect".)

5 AMT Lessons

Amazon's Mechanical Turk and other crowd-sourcing services are widely used in NLP to collect data (Snow et al., 2008), with guides available that provide advice on how best to make use of such services (Callison-Burch and Dredze, 2010). Whether engaging with crowd-sourcing services such as AMT as a requester or worker, however, there is some degree of risk, primarily because of the anonymity that is assured by the services. The requester, in providing payment for potentially large volumes of work, is vulnerable to sub-standard or even robotically completed HITs. In this regard there is a clear sense of "buyer beware" that is part and parcel of using crowd-sourcing services. The worker, on the other hand, earns a relatively low hourly rate, and faces an ongoing risk of having completed HITs declined and of not being reimbursed for diligently completed work. Recently developed online tools provide slightly more power to workers, by enabling requester reviewing and hence allowing workers to identify requesters who too readily reject completed HITs (Irani and Silberman, 2013). And even when workers are paid, rather than volunteers, payment rates are well below the minimum wages that apply in most developed countries (Fort et al., 2011).

Human Ethics Posting HITs on a service such as AMT amounts to research involving humans, and human ethics potentially becomes a concern (Gilles et al., 2011; Fort et al., 2011). Research institutes tend to evolve their own specific human ethics policies for crowd-sourcing tasks. In our particular institution, a two-stage procedure for human ethics approval is in place. An initial stage involves consultation with an advisory group, which functions as a filtering mechanism to determine which applications involving humans need to go through the full ethics application. Our intention to post HITs on AMT was approved at this stage, since material and information collected would not be specifically *about* the subjects.[1] That is, asking AMT workers to assess translations was deemed by the ethics advisory group as research akin to taste-tests or similar market research.

Social Responsibility Besides the issue of personal information, there is an additional ethical concern with regard to payment of workers that remains unresolved in the research community. In non-crowd-sourced research, reimbursing volunteers for work with a small monetary or in-kind reward is common practice and in general is considered ethical. In these experiments the subjects are regarded as volunteers, and the gift or reimbursement is regarded as a gesture of appreciation rather than as payment. With an online service such as AMT, however, the population is mixed: some Turkers may indeed be genuine volunteers, pleased to be able to assist with a research project; and others may be students donating their free time. But almost certainly there are participants – perhaps from developing countries – who rely on their payments as part of their income stream.

It is a human ethics concern if there are large numbers of workers that fall into this last category. Efforts have been made to acquire data about demographics and employment status of workers (Ross et al., 2010; Silberman et al., 2010; Gilles et al., 2011), but little if any of this information is verifiable – in a particularly ironic note, position papers that articulate anti-crowd-sourcing opinions sometimes cite demographics collected through crowd-sourcing as evidence that crowd-sourcing to create datasets is unethical. The service provider itself is probably the only reliable source of information about workers, and even then, there is much that can be hidden behind the screen of Internet anonymity. It can also be argued that, however low the pay rates are compared to minimum wage rates in the country in which the crowd-sourced data is being consumed, to the people carrying out the work, it is better than nothing, and is done voluntarily after full disclosure. Similarly, users of services like AMT observe that if minimum pay rates were to be made compulsory, many of the tasks distributed via crowd-sourcing services would simply be withdrawn, eroding even that modest source of income.

Opportunistic Workers Another issue that arises with crowd-sourcing to create datasets in this regard is that the cloak of anonymity means that there is clear potential for opportunistic workers to attempt to "earn" the payment without doing the work that is required. In some of the literature these workers are referred to as "cheats" (see, for example, Eickhoff and de Vries (2013)) but the reality is that in placing HITs we are seeking to get judgments completed spending as little money as possible; and from the point of view of the workers, their objective is to earn the revenue associated with each HIT spending as little time as possible. That is, both parties to the transaction are seeking to maximize their return. Hence, rather than calling them cheats, we prefer to refer to such workers as being *opportunistic*, or as being *aggressive optimizers*.

Amazon provides some built-in mechanisms to protect requesters from opportunistic workers, and in initial trial HITs we tried some of these restrictions, ultimately retaining some and dropping others. We started with the most conservative restrictions in an attempt to get the best quality data, and applied a *location restriction* according to the target language, in a quest to get native speakers performing the evaluations. We also made use of the *master workers* restriction, which limits workers to a special subset of known (to AMT) high quality workers, at the cost of a slightly higher AMT administration fee. When we applied these restrictions, the response rate was, however, extremely low – possibly due to the combination of the restrictions with too low a payment level. We then reduced the worker restriction from master worker to a 95% previous HIT approval rating. This resulted in a dramatic increase in the response rate for English HITs, but the response

[1] AnonInstitute ethics application reference number 1238934.

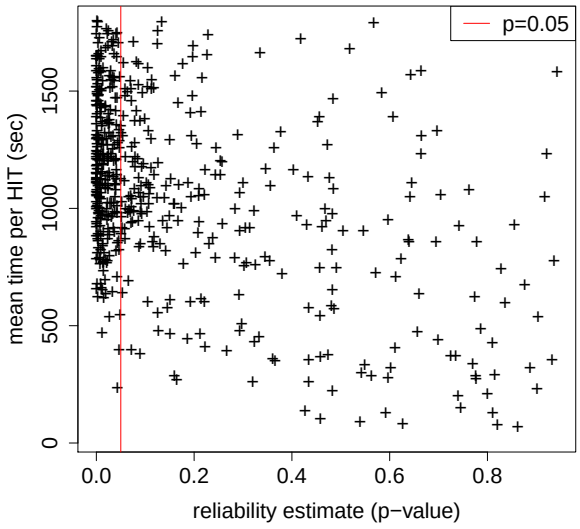

Figure 4: Mean time per 100-translation HIT plotted against workers reliability estimate p-value (lower p-values signify more reliable workers).

rates for the other HITs (restricted to France, Germany, Spain and the Czech Republic, respectively) remained very low. We therefore removed the location restriction for these four languages. The English HITs were location-restricted to US residents throughout the collection process.

The slightly unusual structure of our HITs (each contained 100 judgments) exacerbated the difficulty of deciding what a payment should be. For a HIT of 100 fluency judgments, we set payment at US$0.50. Based on the time that workers took to complete each hit, this amounted on average to an hourly rate of US$1.86 when we include all workers, and US$1.61 for workers who met the quality control threshold (described below).

All but the Czech HITs then proceeded with reasonable response rates. At this level of payment there did not appear to be any group of Czech speakers willing to carry out HITs, and ultimately we dropped the English-Czech language pair from the data collection.

Quality Control One set of opportunistic workers were clearly identifiable due to the unusual structure of our HITs – 100 translations each. The time taken for each HIT ranged from 22 seconds to 1,798 seconds (around 30 minutes). It seems highly unlikely that anyone, no matter how expert, could carry out the task of evaluating a translation on average in 0.22 seconds, and these "workers" made such little effort to pass as human we suspect they may in fact be automated systems. Figure 4 shows for each worker their reliability estimate (as a p value computed over their *bad-reference* pairs, as described in Section 3) versus mean time per HIT (100 translations). Fast HIT completion times almost certainly indicate low quality assessments. For good workers, who met the quality control threshold the average time spent per translation was 10.22s.

Note that the "minimum of 95% approval rating from previous work" requirement was in place throughout our experimentation, including the data plotted in Figure 4. The high number of aggressive optimizers we identified reveals the danger of relying solely on a high previous approval rating. One way in which a worker might manipulate their approval rating is by completing HITs that pay no fee. Presumably, approval of no-fee HITs still results in an increase in a worker's approval rating, and requesters are likely to be less diligent when there is no payment at stake.

Lengthy completion times cannot be used as evidence for good quality work, since no information is available as to what a worker was doing between the time they accepted a HIT and when they submitted it. That is, the workers in the top-right corner of the graph are likely to be a mix of people who sought to obscure their lack of effort by delaying their HIT submission, and people who genuinely spent time on the task, but did not have the necessary knowledge to complete it accurately. Fortunately, a reasonable fraction of workers did meet the quality control threshold of $p < 0.05$.

To avoid rejecting HITs completed by genuine (that is, non opportunistic) workers who were not skilled enough to do the task, we did not decline payments solely on the basis of having a high p-value. Instead, we identified obvious random clickers on the basis of mean scores for *bad-reference* items, for system outputs, and for reference translations. Table 4 shows typical data for the three facets, with worker *D* suspected of being an aggressive optimizer. The HITs from such workers were rejected, and payments declined.

Data Collected Overall, a total of 536 workers generated a total of 91,100 fluency judgments including repeated items. Of these, 49% of workers reached the quality control threshold; they accounted for 57% of the HITs. Four workers com-

Worker	A	B	C	*D*	E	F	G	H	I	J	K	L	M	N
bad-reference	37.7	31.5	19.4	*87.6*	8.5	8.1	3.0	6.6	20.2	7.4	14.2	57.0	50.4	29.3
system	46.5	52.8	41.8	*85.2*	47.0	35.4	16.0	38.0	31.6	33.1	42.5	59.1	66.0	52.7
reference	64.1	92.6	88.8	*81.3*	53.7	42.7	89.7	76.8	92.5	82.4	74.8	60.7	83.9	59.2

Table 4: Mean scores judged by fourteen workers for *bad-reference* items, for system outputs, and for reference translations. Worker *D*'s behavior is sufficiently anomalous that their HITs were rejected.

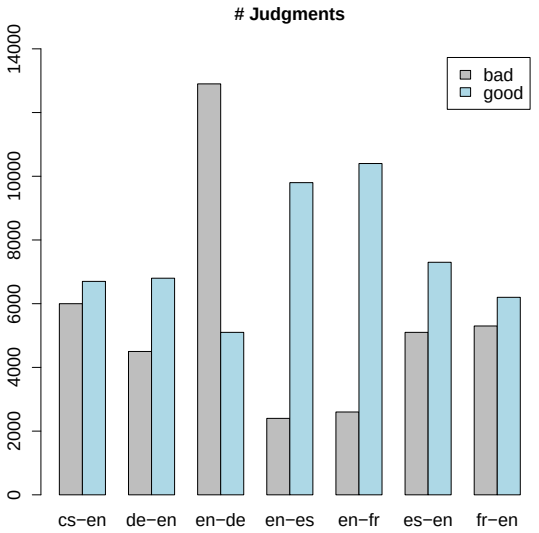

Figure 5: Numbers of judgments by language pair, categorized by whether they reached the desired quality level for *bad-reference* items.

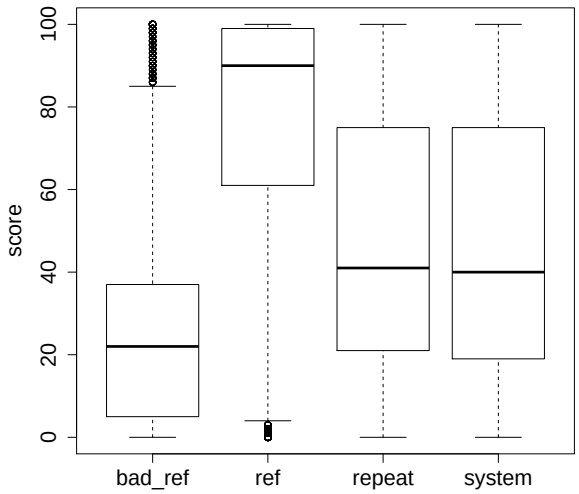

Figure 6: Scores of good workers for (left to right) *bad-reference* degraded translations; reference translations; *ask_again* translations; and normal system outputs.

pleted more than 10 HITs; and one worker completed 50 HITs. Figure 5 shows how the balance between good-quality and bad-quality judgments varied across target languages, with numbers of good French and Spanish judgments far exceeding those of both English and German, and a majority of workers who completed the German task not reaching the quality control threshold. German HITs had a slower response rate, probably due to fewer AMT workers being speakers of German than French, Spanish or English. In total, 28 of the 536 workers had an average HIT completion time of less than 5 minutes, and 17 of those were for German HITs. In addition 3 workers completed HITs for more than one target language; since we had requested native speakers, that was also regarded as being grounds for rejection. The German HITs were targeted by opportunistic workers, but it is interesting that the seemingly equally-tempting Czech HITs were not.

Figure 6 shows the score distributions of the good workers over the four types of item in each HIT, and confirms that the categorization of workers into good and bad yielded the desired outcome.

6 Conclusion

Human evaluation forms the basis upon which all empirical machine translation research is founded, whether it be directly through employing humans to assess the quality of machine translation output or through the use of automatic metrics that have been validated by correlation with human judgments. We have collected a large dataset of human assessments of machine translation system outputs, employing a range of different assessment configurations. This data set will be made public once it has been fully collated and meta-data added to it, and will form a resource for further evaluation of machine translation research.

Acknowledgments This work was supported by the Australian Research Council. Ethics application (number 1238934) submitted November 2012, with outcome of "no approval needed due to the impersonal nature of the material".

References

A. Belz and E. Kow. 2011. Discrete vs. continuous rating scales for language evaluation in NLP. In *Proc. 49th Ann. Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 230–235, Portland, USA.

O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proc. 8th Wkshp. on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria.

C. Callison-Burch and M. Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proc. NAACL HLT 2010 Wkshp. on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles, USA.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proc. 2nd Wkshp. Statistical Machine Translation*, pages 136–158, Prague, Czech Republic.

C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proc. 7th Wkshp. Statistical Machine Translation*, pages 10–51, Montreal, Canada.

C. Eickhoff and A. P. de Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, 16(2):121–137.

K. Fort, G. Adda, and K. B. Cohen. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

A. Gilles, B. Sagot, K. Fort, and J. Mariani. 2011. Crowdsourcing for language resource development: Critical analysis of Amazon Mechanical Turk overpowering use. In *Proc. 5th Language and Technology Conf.*, Poznań, Poland.

Y. Graham, T. Baldwin, A. Harwood, A. Moffat, and J. Zobel. 2012. Measurement of progress in machine translation. In *Proc. Australasian Language Technology Wkshp.*, pages 70–78, Dunedin, New Zealand.

Y. Graham, T. Baldwin, A. Moffat, and J. Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proc. 7th Linguistic Annotation Wkshp.*, pages 33–41, Sofia, Bulgaria.

L. Irani and M. S. Silberman. 2013. Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. In *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems*, pages 611–620, Paris, France.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research, Thomas J. Watson Research Center.

J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. 2010. Who are the crowdworkers? Shifting demographics in Mechanical Turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 2863–2872.

M. S. Silberman, J. Ross, L. Irani, and B. Tomlinson. 2010. Sellers' problems in human computation markets. In *Proc. ACM SIGKDD Wkshp. on Human Computation*, pages 18–21, Washington DC, USA.

R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and fast – But is it good? Evaluating non-expert annotations for natural language tasks. In *Proc. Conf. Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, USA.