Continuous Measurement Scales in Human Evaluation of Machine Translation

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

We explore the use of continuous rating scales for human evaluation in the context of machine translation evaluation, comparing two assessor-intrinsic quality-control techniques that do not rely on agreement with expert judgments. Experiments employing Amazon's Mechanical Turk service show that quality-control techniques made possible by the use of the continuous scale show dramatic improvements to intra-annotator agreement of up to $+0.101$ in the kappa coefficient, with inter-annotator agreement increasing by up to $+0.144$ when additional standardization of scores is applied.

1 Introduction

Human annotations of language are often required in natural language processing (NLP) tasks for evaluation purposes, in order to estimate how well a given system mimics activities traditionally performed by humans. In tasks such as machine translation (MT) and natural language generation, the system output is a fully-formed string in a target language. Annotations can take the form of direct estimates of the quality of those outputs or be structured as the simpler task of ranking competing outputs from best-to-worst (Callison-Burch et al., 2012).

A direct estimation method of assessment, as opposed to ranking outputs from best-to-worst, has the advantage that it includes in annotations not only that one output is better than another, but also the degree to which that output was better than the other. In addition, direct estimation of quality within the context of machine translation extends the usefulness of the annotated data to other tasks such as quality-estimation (Callison-Burch et al., 2012).

For an evaluation to be credible, the annotations must be credible. The simplest way of establishing this is to have the same data point annotated by multiple annotators, and measure the agreement between them. There has been a worrying trend in recent MT shared tasks – whether the evaluation was structured as ranking translations from best-to-worst, or by direct estimation of fluency and adequacy – of agreement between annotators decreasing (Callison-Burch et al., 2008; Callison-Burch et al., 2009; Callison-Burch et al., 2010; Callison-Burch et al., 2011; Callison-Burch et al., 2012). Inconsistency in human evaluation of machine translation calls into question conclusions drawn from those assessments, and is the target of this paper: by revising the annotation process, can we improve annotator agreement, and hence the quality of human annotations?

Direct estimates of quality are intrinsically continuous in nature, but are often collected using an interval-level scale with a relatively low number of categories, perhaps to make the task cognitively easier for human assessors. In MT evaluation, five and seven-point interval-level scales are common (Callison-Burch et al., 2007; Denkowski and Lavie, 2010). However, the interval-level scale commonly used for direct estimation of translation quality (and other NLP annotation tasks) forces human judges to discretize their assessments into a fixed number of categories, and this process could be a cause of inconsistency in human judgments. In particular, an assessor may be repeatedly forced to choose between two categories, neither of which really fits their judgment. The continuous nature of translation quality assessment, as well as the fact that many statistical methods exist that can be applied to continuous data but not interval-level data, motivates our trial of a continuous rating scale.

We use human judgments of translation fluency as a test case and compare consistency levels when

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 33–41,
Sofia, Bulgaria, August 8-9, 2013. ©2013 Association for Computational Linguistics

the conventional 5-point interval-level scale and a continuous visual analog scale (VAS) are used for human evaluation. We collected data via Amazon's Mechanical Turk, where the quality of annotations is known to vary considerably (Callison-Burch et al., 2010). As such, we test two quality-control techniques based on statistical significance – made possible by the use of the continuous rating scale – to intrinsically assess the quality of individual human judges. The quality-control techniques are not restricted to fluency judgments and are relevant to more general MT evaluation, as well as other NLP annotation tasks.

2 Machine Translation Fluency

Measurement of fluency as a component of MT evaluation has been carried out for a number of years (LDC, 2005), but it has proven difficult to acquire consistent judgments, even from expert assessors. Evaluation rounds such as the annual Workshop on Statistical Machine Translation (WMT) use human judgments of translation quality to produce official rankings in shared tasks, initially using an two-item assessment of fluency and adequacy as separate attributes, and more recently by asking judges to simply rank system outputs against one another according to "which translation is better". However, the latter method also reports low levels of agreement between judges. For example, the 2007 WMT reported low levels of consistency in fluency judgments in terms of both intra-annotator agreement (intra-aa), with a kappa coefficient of $\kappa = 0.54$ (moderate), and inter-annotator agreement (inter-aa), with $\kappa = 0.25$ (slight). Adequacy judgments for the same data received even lower scores: $\kappa = 0.47$ for intra-aa, and $\kappa = 0.23$ for inter-aa.

While concerns over annotator agreement have seen recent WMT evaluations move away from using fluency as an evaluation component, there can be no question that fluency is a useful means of evaluating translation output. In particular, it is not biased by reference translations. The use of automatic metrics is often criticized by the fact that a system that produces a good translation which happens not to be similar to the reference translations will be unfairly penalized. Similarly, if human annotators are provided with one or more reference sentences, they may inadvertently favor translations that are similar to those references. If fluency is judged independently of adequacy, no reference translation is needed, and the bias is removed.

In earlier work, we consider the possibility that *translation quality* is a hypothetical construct (Graham et al., 2012), and suggest applying methods of validating measurement of psychological constructs to the validation of measurements of translation quality. In psychology, a scale that employs more items as opposed to fewer is considered more valid. Under this criteria, a two-item (fluency and adequacy) scale is more valid than a single-item translation quality measure.

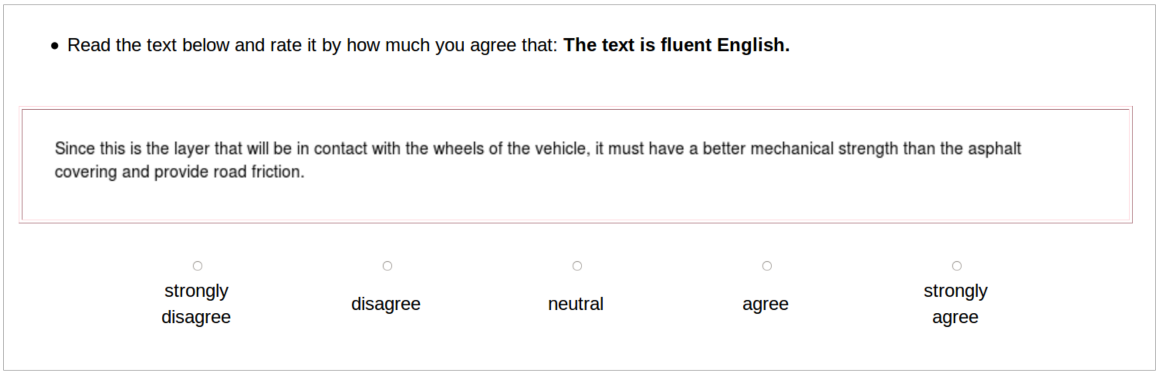
3 Measurement Scales

Direct estimation methods are designed to elicit from the subject a direct quantitative estimate of the magnitude of an attribute (Streiner and Norman, 1989). We compare judgments collected on a visual analog scale (VAS) to those using an interval-level scale presented to the human judge as a sequence of radio-buttons. The VAS was first used in psychology in the 1920's, and prior to the digital age, scales used a line of fixed length (usually 100mm in length), with anchor labels at both ends, and to be marked by hand with an "X" at the desired location (Streiner and Norman, 1989).

When an interval-scale is used in NLP evaluation or other annotation tasks, it is commonly presented in the form of an adjectival scale, where categories are labeled in increasing/decreasing quality. For example, an MT evaluation of fluency might specify 5 = "Flawless English", 4 = "Good English", 3 = "Non-native English", 2 = "Disfluent English", and 1 = "Incomprehensible" (Callison-Burch et al., 2007; Denkowski and Lavie, 2010).

With both a VAS and an adjectival scale, the choice of labels can be critical. In medical research, patients' ratings of their own health have been shown to be highly dependent on the exact wording of descriptors (Seymour et al., 1985). Alexandrov (2010) provides a summary of the extensive literature on the numerous issues associated with adjectival scale labels, including bias resulting from positively and negatively worded items not being true opposites of one another, and items intended to have neutral intensity in fact proving to have unique conceptual meanings.

Likert scales avoid the problems associated with adjectival labels, by structuring the question as a simple statement that the respondent registers their level of (dis)agreement with. Figure 1 shows



- Read the text below and rate it by how much you agree that: **The text is fluent English.**

Since this is the layer that will be in contact with the wheels of the vehicle, it must have a better mechanical strength than the asphalt covering and provide road friction.

Figure 1: Amazon Mechanical Turk interface for fluency judgments with a Likert-type scale.

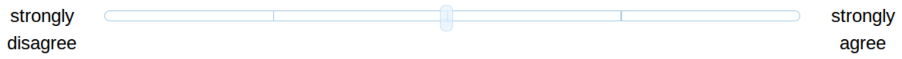


Figure 2: Continuous rating scale for fluency judgments with two anchors.

the Likert-type interval-level scale we use to collect fluency judgments of MT output, and Figure 2 shows an equivalent VAS using the two most extreme anchor labels, *strongly disagree* and *strongly agree*.

4 Crowd-sourcing Judgments

The volume of judgments required for evaluation of NLP tasks can be large, and employing experts to undertake those judgments may not always be feasible. Crowd-sourcing services via the Web offer an attractive alternative, and have been used in conjunction with a range of NLP evaluation and annotation tasks. Several guides exist for instructing researchers from various backgrounds on using Amazon's Mechanical Turk (AMT) (Gibson et al., 2011; Callison-Burch, 2009), and allowance for the use of AMT is increasingly being made in research grant applications, as a cost-effective way of gathering data. Issues remain in connection with low payment levels (Fort et al., 2011); nevertheless, Ethics Approval Boards are typically disinterested in projects that make use of AMT, regarding AMT as being a purchased service rather than a part of the experimentation that may affect human subjects.

The use of crowd-sourced judgments does, however, introduce the possibility of increased inconsistency, with service requesters typically having no specific or verifiable knowledge about any given worker. Hence, the possibility that a worker is acting in good faith but not performing the task well must be allowed for, as must the likelihood that some workers will quite ruthlessly seek to minimize the time spent on the task, by deliberately giving low-quality or fake answers. Some workers may even attempt to implement automated responses, so that they get paid without having to do the work they are being paid for.

For example, if the task at hand is that of assessing the fluency of text snippets, it is desirable to employ native speakers. With AMT the requester has the ability to restrict responses to only workers who have a specified skill. But that facility does not necessarily lead to confidence – there is nothing stopping a worker employing someone else to do the test for them. Devising a test that reliably evaluates whether or not someone is a native speaker is also not at all straightforward.

Amazon allow location restrictions, based on the registered residential address of the Turker, which can be used to select in favor of those likely to have at least some level of fluency (Callison-Burch et al., 2010). We initially applied this restriction to both sets of judgments in experiments, setting the task up so that only workers registered in Germany could evaluate the to-German translations, for example. However, very low response rates for languages other than to-English were problematic, and we also received a number of apparently-genuine requests from native speakers residing outside the target countries. As a result, we removed all location restrictions other than for the to-English tasks.[1]

Crowd-sourcing judgments has the obvious risk of being vulnerable to manipulation. On the other hand, crowd-sourced judgments also offer the potential of being *more* valid than those of experts, since person-in-the-street abilities might be a more useful yardstick for some tasks than informed academic judgment, and because a greater number of judges may be available.

Having the ability to somehow evaluate the quality of the work undertaken by a Turker is thus highly desirable. We would like to be able to put in place a mechanism that filters out non-native speakers; native speakers with low literacy levels; cheats; and robotic cheats. That goal is considered in the next section.

5 Judge-Intrinsic Quality Control

One common method of quality assessment for a new process is to identify a set of "gold-standard" items that have been judged by experts and whose merits are agreed, present them to the new process or assessor, and then assess the degree to which the new process and the experts "agree" on the outcomes (Snow et al., 2008; Callison-Burch et al., 2010). A possible concern is that even experts can be expected to disagree (and hence have low inter-aa levels), meaning that disagreement with the new process will also occur, even if the new process is a reliable one. In addition, the quality of the judgments collected is also assessed via agreement levels, meaning that any filtering based on a quality-control measure that uses agreement will automatically increase consistency, even to the extent of recalibrating non-expert workers' responses to more closely match expert judgments (Snow et al., 2008). Moreover, if an interval-level scale is used, standardized scores cannot be employed, so a non-expert who is more lenient than the experts, but in a reliable and systematic manner, might still have their assessments discarded.

For judgments collected on a continuous scale, statistical tests based on difference of means (over assessors) are possible. We structure our human intelligence tasks (HITs) on Mechanical Turk in groups of 100 in a way that allows us to control assignment of repeat item pairs to workers, so that statistical tests can later be applied to an individual worker's score distributions for repeat items. Workers were made aware of the task structure before accepting it – the task preview included a message *This HIT consists of 100 fluency assessments, you have 0 so far complete*.

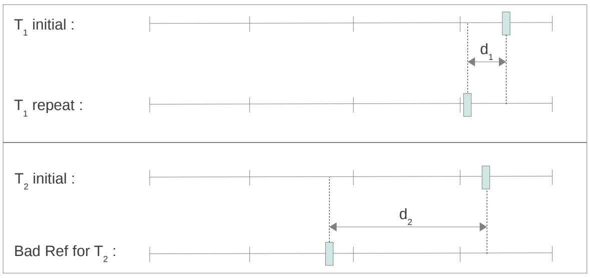


Figure 3: Intrinsic quality-control distributions for an individual judge.

We refer to the repeat items in a HIT as *ask_again* translations. In addition, we inserted a number of *bad_reference* pairs into each HIT, with a *bad_reference* pair consisting of a genuine MT system output, and a distorted sentence derived from it, expecting that its fluency was markedly worse than that of the corresponding system output. This was done by randomly selecting two words in the sentence and duplicating them in random locations not adjacent to the original word and not in the initial or sentence-final position. Any other degradation method could also be used, so long as it has a high probability of reducing the fluency of the text, and provided that it is not immediately obvious to the judges.

Insertion of *ask_again* and *bad_reference* pairs into the HITs allowed two measurements to be made for each worker: when presented with an *ask_again* pair, we expect a conscientious judge to give similar scores (but when using a continuous scale, certainly not identical), and on *bad_reference* pairings a conscientious judge should reliably give the altered sentence a lower score. The wide separation of the two appearances of an *ask_again* pair makes it unlikely that a judge would remember either the sentence or their first reaction to it, and backwards movement through the sentences comprising each HIT was not possible. In total, each HIT contained 100 sen-

[1] It has also been suggested that AMT restricts Turker registration by country; official information is unclear about this.

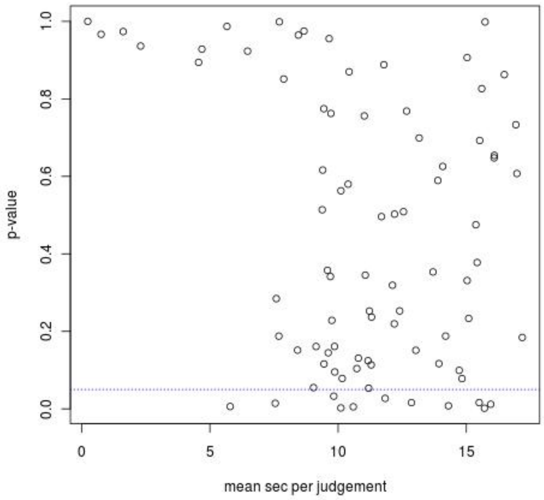


Figure 4: Welch's t-test reliability estimates plotted against mean seconds per judgment.

tences, including 10 *bad_reference* pairs, and 10 *ask_again* pairs.

Figure 3 illustrates these two types of pairs, presuming that over the course of one or more HITs each worker has assessed multiple *ask_again* pairs generating the distribution indicated by d_1, and also multiple *bad_reference* pairs, generating the distribution indicated by d_2. As an estimate of the reliability of each individual judge we apply a t-test to compare *ask_again* differences with *bad_reference* differences, with the expectation that for a conscientious worker the latter should be larger than the former. Since there is no guarantee that the two distributions of d_1 and d_2 have the same variance, we apply Welch's adaptation of the Student t-test.

The null hypothesis to be tested for each AMT worker is that the score difference for *ask_again* pairs is not less than the score difference for *bad_reference* pairs. Lower p values mean more reliable workers; in the experiments that are reported shortly, we use $p < 0.05$ as a threshold of reliability. We also applied the non-parametric Mann-Whitney test to the same data, for the purpose of comparison, since there is no guarantee that d_1 and d_2 will be normally distributed for a given assessor.

The next section provides details of the experimental structure, and then describes the outcomes in terms of their effect on overall system rankings. As a preliminary indication of Turker behavior, Figure 4 summarizes some of the data that was obtained. Each plotted point represents one AMT worker who took part in our experiments, and the horizontal axis reflects their average per-judgment time (noting that this is an imprecise measurement, since they may have taken phone calls or answered email while working through a HIT, or simply left the task idle to help obscure a lack of effort). The vertical scale is the p value obtained for that worker when the *ask_again* distribution is compared to their *bad_reference* distribution, with a line at $p = 0.05$ indicating the upper limit of the zone for which we are confident that they had a different overall response to *ask_again* pairs than they did to *bad_reference* pairs. Note the small number of very fast, very inaccurate workers at the top left; we have no hesitation in calling them unconscientious (and declining to pay them for their completed HITs). Note also the very small number of workers for which it was possible to reliably distinguish their *ask_again* behavior from their *bad_reference* behavior.

6 Experiments

HIT Structure

A sample of 560 translations was selected at random from the WMT 2012 published shared task dataset for a range of language pairs, with segments consisting of 70 translations, each assigned to a total of eight distinct HITs. The sentences were generated as image files, as recommended for judgment of translations (Callison-Burch, 2009). Each HIT was presented to a worker as a set of 100 sentences including a total of 30 quality control items, with only one sentence visible on-screen at any given time. Each quality control item comprised a pair of corresponding translations, widely separated within the HIT. Three kinds of quality control pairs were used:

- *ask_again*: system output and exact repeat;
- *bad_reference*: system output and an altered version of it with noticeably lower fluency; and
- *good_reference*: system output and the corresponding human produced reference translation (as provided in the released WMT data).

Each HIT consisted of 10 groups, each containing 10 sentences: 7 "normal" translations, plus one of each type of quality control translation drawn from one of the other groups in the HIT in such a way that 40–60 judgments would be completed between the elements of any quality-control pair.

Consistency of Human Judgments

Using judgments collected on the continuous rating scale, we first examine assessor consistency based on Welch's t-test and the non-parametric Mann-Whitney U-test. In order to examine the degree to which human assessors assign consistent scores, we compute mean values of d_1 (Figure 3) when *ask_again* pairs are given to the same judge, and across pairs of judges. Three sets of results are shown: the raw unfiltered data; data filtered according to $p < 0.05$ according to the quality-control regime described in the previous section using the Welch's t-test; and data filtered using the Mann-Whitney U-test. Table 1 shows that the t-test indicates that only 13.1% of assessors meet quality control hurdle, while a higher proportion, 35.7%, of assessors are deemed acceptable.

The stricter filter, Welch's t-test, yields more consistent scores for same-judge repeat items: decreases of 4.5 (mean) and 4.2 (sd) are observed when quality control is applied. In addition, results for Welch's t-test show high levels of consistency for same-judge repeat items: an average difference of only 9.5 is observed, which is not unreasonable, given that the scale is 100 points in length and a 10-point difference corresponds to just 60 pixels on the screen.

For repeat items rated by distinct judges, both filtering methods decrease the mean difference in scores compared to the unfiltered baseline, with the two tests giving similar improvements.

When an interval-level scale is used to evaluate the data, the Kappa coefficient is commonly used to evaluate consistency levels of human judges (Callison-Burch et al., 2007), where $\Pr(a)$ is the relative observed agreement among raters, and $\Pr(e)$ is the hypothetical probability of chance agreement:

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

In order to use the Kappa coefficient to compare agreement levels for the interval-level and continuous scales, we convert continuous scale scores to a target number of interval categories. We do this primarily for a target number of five, as this best provides a comparison between scores for the 5-point interval-level scale. But we also present results for targets of four and two categories, since the continuous scale is marked at the midway and quarter points, providing implicit intervals. A two-category is also interesting if the assessment process is regarded as dichotomizing to only include for each translation whether or not the judge considered it to be "good" or "bad". Use of statistical difference of means tests on interval-level data is not recommended; but for the purpose of illustration, we also applied Welch's t-test to quality control workers that completed the interval-level HITs, with the same threshold of $p < 0.05$.

Tables 2 and 3 show intra-annotator agreement for the five-point interval scale and continuous scales, with and without quality control.[2] Results for repeat items on the interval-level scale show that quality control only alters intra-aa marginally ($\Pr(a)$ increases by 1%), and that inter-aa levels worsen ($\Pr(a)$ decreases by 6.2%). This confirms that applying statistical tests to interval-level data is not a suitable way of filtering out low quality workers.

When comparing consistency levels of assessors using the interval-level scale to those of the continuous scale, we observe marginally lower κ coefficients for both intra-aa (-0.009) and inter-aa (-0.041) for the continuous scale. However, this is likely to be in part due to the fact that the continuous scale corresponds more intuitively to 4 categories, and agreement levels for the unfiltered 4-category continuous scale are higher than those collected on the interval-level scale by $+0.023$ intra-aa and $+0.014$ inter-aa.

Applying quality-control on the continuous scale results in dramatic increases in intra-aa levels: $+0.152$ for 5-categories (5-cat), $+0.100$ for 4-categories (4-cat) and $+0.096$ for 2-categories (2-cat). When considering inter-aa levels, quality-control does not directly result in as dramatic an increase, as inter-aa levels increase by $+0.010$ for 5-cat, $+0.006$ for 4-cat and $+0.004$ for 2-cat. It is likely, however, that apparent disagreement between assessors might be due to different assessors judging fluency generally worse or better than one another. The continuous scale allows for scores to be standardized by normalizing scores with respect to the mean and standard deviation of all scores assigned by a given individual judge. We therefore transform scores of each judge into

[2] Note that the mapping from continuous scores to categories was not applied for quality control.

			same judge		distinct judges	
	workers	judgments	mean	sd	mean	sd
Unfiltered	100.0%	100.0%	14.0	18.4	28.9	23.5
Welch's t-test	13.1%	23.5%	9.5	14.2	25.2	21.0
Mann-Whitney U-test	35.7%	48.8%	13.1	17.7	25.0	22.6

Table 1: Mean and standard deviation of score differences for continuous scale with *ask_again* items within a given judge and across two distinct judges, for no quality control (unfiltered), Welch's t-test and Mann-Whitney U-test with a quality-control threshold of $p < 0.05$.

# categ- ories	5-pt. interval unfiltered		5-pt. interval filtered		continuous unfiltered		continuous filtered	
	$\Pr(a)$	κ	$\Pr(a)$	κ	$\Pr(a)$	κ	$\Pr(a)$	κ
5	60.4%	0.505	61.4%	0.517	59.7%	0.496	71.8%	0.647
4	-	-	-	-	64.6%	0.528	72.1%	0.629
2	-	-	-	-	85.2%	0.704	90.0%	0.800

Table 2: Intra-annotator (same judge) agreement levels for 5-point interval and continuous scales for unfiltered judgments and judgments of workers with $p < 0.05$ for Welch's t-test.

corresponding z-scores and use percentiles of the combined set of all scores to map z-scores to categories where a score falling in the bottom 20 th percentile corresponds to *strongly disagree*, scores between the 20 th and 40 th percentile to *disagree*, and so on. Although this method of transformation is somewhat harsh on the continuous scale, since scores no longer correspond to different locations on the original scale, it nevertheless shows an increase in consistency of $+0.05$ (5-cat), $+0.086$ (4-cat) and $+0.144$ (2-cat). However, caution must be taken when interpreting consistency for standardized scores, as can be seen from the increase in agreement observed when unfiltered scores are standardized.

Table 4 shows a breakdown by target language of the proportion of judgments collected whose scores met the significance threshold of $p < 0.05$. Results appear at first to have shockingly low levels of high quality work, especially for English and German. When running the tasks in Mechanical Turk, it is worth noting that we did not adopt statistical tests to automatically accept/reject HITs and we believe this would be rather harsh on workers. Our method of quality control is a high bar to reach and it is likely that many workers that do not meet the significance threshold would still have been working in good faith. In practice, we individually examined mean scores for reference translation, system outputs and *bad_reference* pairs, and only declined payment when there was no doubt the response was either automatic or extremely careless.

English	German	French	Spanish
10.0%	0%	57.9%	62.5%

Table 4: High quality judgments, by language.

The structure of the task and the fact that the quality-control items were somewhat hidden may have lulled workers into a false sense of complacency, and perhaps encouraged careless responses. However, even taking this into consideration, the fact that *none* of the German speaking assessors and just 10% of English speaking assessors reached our standards serves to highlight the importance of good quality-control techniques when employing services like AMT. In addition, the risk of getting low quality work for some languages might be more risky than for others. The response rate for high quality work for Spanish and French was so much higher than German and English, perhaps by chance, or perhaps the result of factors that will be revealed in future experimentation.

System Rankings

As an example of the degree to which system rankings are affected by applying quality control, for the language direction for which we achieved the highest number of high quality assessments, English-to-Spanish, we include system rankings by mean score with each measurement scale, with and without quality control and for mean z-scores

#	5-pt. interval		5-pt. interval		continuous		continuous		cont. standrdzed.		cont. standrdzed.	
categ-	unfiltered		qual.-controlled		unfiltered		qual.-controlled		unfiltered		qual.-controlled	
ories	Pr(a)	κ	Pr(a)	κ	Pr(a)	κ	Pr(a)	κ	Pr(a)	κ	Pr(a)	κ
5	33.0%	0.16	26.8%	0.084	29.5%	0.119	30.3%	0.128	30.2%	0.1272	33.5%	0.169
4	-	-	-	-	38.1%	0.174	38.5%	0.180	35.5%	0.1403	44.5%	0.260
2	-	-	-	-	66.5%	0.331	66.8%	0.335	75.5%	0.5097	73.8%	0.475

Table 3: Inter-annotator (distinct judge) agreement levels for 5-point interval and continuous scales for unfiltered judgments and judgments of workers with $p < 0.05$ for Welch's t-test.

								z-scores	
5-pt.		5-pt.		continuous		continuous		continuous	
unfiltered		qual.-controlled		unfiltered		qual.-controlled		qual.-controlled	
Sys A	2.00	Sys A	2.00	Sys E	69.60	Sys E	74.39	Sys E	0.43
Sys B	1.98	Sys D	1.97	Sys B	61.78	Sys F	65.07	Sys B	0.16
Sys C	1.98	Sys F	1.95	Sys G	60.21	Sys G	64.51	Sys G	0.08
Sys D	1.98	Sys C	1.95	Sys F	59.38	Sys B	63.68	Sys D	0.06
Sys E	1.98	Sys E	1.95	Sys D	59.05	Sys D	63.52	Sys C	0.02
Sys F	1.97	Sys B	1.94	Sys A	57.44	Sys C	61.33	Sys F	0.01
Sys G	1.97	Sys G	1.93	Sys I	56.31	Sys A	58.43	Sys H	–0.03
Sys H	1.96	Sys H	1.90	Sys C	55.82	Sys I	57.46	Sys I	–0.07
Sys I	1.96	Sys I	1.88	Sys H	55.27	Sys H	57.04	Sys A	–0.10
Sys J	1.94	Sys J	1.81	Sys J	50.46	Sys J	50.73	Sys J	–0.23
Sys K	1.90	Sys K	1.76	Sys K	44.62	Sys K	41.25	Sys K	–0.47

Table 5: WMT system rankings based on approximately 80 randomly-selected fluency judgments per system, with and without quality control for radio button and continuous input types, based on German-English. The quality control method applied is annotators who score worsened system output and genuine system outputs with statistically significant lower scores according to paired Student's t-test.

when raw scores are normalized by individual assessor mean and standard deviation. The results are shown in Table 5. (Note that we do not claim that these rankings are indicative of actual system rankings, as only fluency of translations was assessed, using an average of just 55 translations per system.)

When comparing system rankings for unfiltered versus quality-controlled continuous scales, firstly the overall difference in ranking is not as dramatic as one might expect, as many systems retain the same rank order, with only a small number of systems changing position. This happens because random-clickers cannot systematically favor any system, and positive and negative random scores tend to cancel each other out. However, even having two systems ordered incorrectly is of concern; careful quality control, and the use of normalization of assessors' scores may lead to more consistent outcomes. We also note that incorrect system orderings may lead to flow-on effects for evaluation of automatic metrics.

The system rankings in Table 5 also show how the use of the continuous scale can be used to rank systems according to z-scores, so that individual assessor preferences over judgments can be ameliorated. Interestingly, the system that scores closest to the mean, Sys F, corresponds to the baseline system for the shared task with a z-score of 0.01.

7 Conclusion

We have compared human assessor consistency levels for judgments collected on a five-point interval-level scale to those collected on a continuous scale, using machine translation fluency as a test case. We described a method for quality-controlling crowd-sourced annotations that results in marked increases in intra-annotator consistency and does not require judges to agree with experts. In addition, the use of a continuous scale allows scores to be standardized to eliminate individual judge preferences, resulting in higher levels of inter-annotator consistency.

Acknowledgments

This work was funded by the Australian Research Council. Ondřej Bojar, Rosa Gog, Simon Gog, Florian Hanke, Maika Vincente Navarro, Pavel Pecina, and Djame Seddah provided translations of task instructions, and feedback on published HITs.

References

A. Alexandrov. 2010. Characteristics of single-item measures in Likert scale format. *The Electronic Journal of Business Research Methods*, 8:1–12.

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, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2008. Further meta-evaluation of machine translation. In *Proc. 3rd Wkshp. Statistical Machine Translation*, pages 70–106, Columbus, Ohio.

C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc. 4th Wkshp. Statistical Machine Translation*, pages 1–28, Athens, Greece.

C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proc. 5th Wkshp. Statistical Machine Translation*, pages 17–53, Uppsala, Sweden.

C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proc. 6th Wkshp. Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland.

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. Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proc. Conf. Empirical Methods in Natural Language Processing*, pages 286–295, Singapore.

M. Denkowski and A. Lavie. 2010. Choosing the right evaluation for machine translation: An examination of annotator and automatic metric performance on human judgement tasks. In *Proc. 9th Conf. Assoc. Machine Translation in the Americas (AMTA)*, Denver, Colorado.

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

E. Gibson, S. Piantadosi, and K. Fedorenko. 2011. Using Mechanical Turk to obtain and analyze English acceptability judgments. *Language and Linguistics Compass*, 5/8:509–524.

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.

LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report, Linguistic Data Consortium. Revision 1.5.

R. A. Seymour, J. M. Simpson, J. E. Charlton, and M. E. Phillips. 1985. An evaluation of length and end-phrase of visiual analogue scales in dental pain. *Pain*, 21:177–185.

R. Snow, B. O'Connor, D. Jursfsky, 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, Hawaii.

D. L. Streiner and G. R. Norman. 1989. *Health Measurement Scales: A Practical Guide to their Development and Use*. Oxford University Press, fourth edition.