Natural Language Engineering **1** (1): 1–28. Printed in the United Kingdom

Can Machine Translation Systems be Evaluated by the Crowd Alone?

Yvette Graham[♠♡] Timothy Baldwin[♠] Alistair Moffat[♠] Justin Zobel[♠]

♠ *Department of Computing and Information Systems*
The University of Melbourne

{tbaldwin|ammoffat|jzobel}@unimelb.edu.au

♡ *ADAPT Centre*
School of Computer Science and Statistics
Trinity College Dublin

graham.yvette@gmail.com

(*Received day month year; revised day month year*)

Abstract

Crowd-sourced assessments of machine translation quality allow evaluations to be carried out cheaply and on a large scale. It is essential, however, that the crowd's work be filtered to avoid contamination of results through the inclusion of false assessments. One method is to filter via agreement with experts, but even amongst experts agreement levels may not be high. In this paper, we present a new methodology for crowd-sourcing human assessments of translation quality, which allows individual workers to develop their own individual assessment strategy. Agreement with experts is no longer required, and a worker is deemed reliable if they are consistent relative to their own previous work. Individual translations are assessed in isolation from all others in the form of direct estimates of translation quality. This allows more meaningful statistics to be computed for systems and enables significance to be determined on smaller sets of assessments. We demonstrate the methodology's feasibility in large-scale human evaluation through replication of the human evaluation component of WMT shared translation task for two language pairs, Spanish-to-English and English-to-Spanish. Results for measurement based solely on crowd-sourced assessments show system rankings in line with those of the original evaluation. Comparison of results produced by the relative preference approach and the direct estimate method described here demonstrate that the direct estimate method has a substantially increased ability to identify significant differences between translation systems.

1 Introduction

The ability to develop and refine machine translation (MT) systems is critically reliant on the availability of reliable methods of assessing the quality of translations. Expert assessment of translation quality is widely held as a 'gold standard' yardstick, but is costly. Automatic evaluation is often used as a substitute, as a means of, for example, supporting automatic MT system comparison, rapid MT deployment, and automatic tuning of system parameters (Och, 2003; Kumar and Byrne, 2004). However, it is well documented that the

Graham, Yvette, Timothy Baldwin, Alistair Moffat and Justin Zobel (to appear) Can Machine Translation Systems be Evaluated by the Crowd Alone? Natural Language Engineering.

automatic MT evaluation metrics in current use fall short of human assessment (Callison-Burch, 2009; Graham, Mathur, and Baldwin, 2014). As a result, manual evaluation is still widely used as the primary means of evaluation in large-scale shared tasks, such as the annual Workshop on Statistical Machine Translation (WMT) (Bojar et al., 2014). Additionally, for the purpose of benchmarking and development of automatic MT evaluation metrics, it is vital that data sets are developed with high-quality human assessments and over a variety of language pairs.

Human assessment can be expert or crowd-sourced, or a mixture of these. To date, WMT shared translation tasks have mainly used expert-only human evaluations as the basis of official evaluation (Callison-Burch, Fordyce, Koehn, Monz, and Schroeder, 2007, 2008; Callison-Burch, Koehn, Monz, and Schroeder, 2009; Callison-Burch et al., 2010; Bojar et al., 2014). There was experimentation with a mix of expert and crowd-sourced judgments to produce official results in 2012–2013 (Callison-Burch et al., 2012; Bojar et al., 2013), but in 2014, the organizers reverted to expert-only assessments due to worryingly low inter-annotator agreement rates for some language pairs.

We argue that it is possible to measure MT systems reliably based on crowd-sourced judgments alone. To demonstrate that claim, we have explored and quantified an approach to crowd-sourcing of human assessments of translation quality using the Amazon Mechanical Turk (AMT) platform. Key features of our proposed methodology are:

- It can be used to assess both adequacy and fluency;
- It requires only monolingual annotators conversant in the target language, thus allowing use of a larger pool of lower-skilled annotators than is possible with standard manual evaluation approaches, which require bilingual annotators with high-level competency in both the source and target languages;
- The ratings are captured via direct estimates on a 100-point Likert scale, enabling fine-grained statistical analysis (Graham, Baldwin, Moffat, and Zobel, 2013);
- It incorporates mechanisms for quality control, based on internal consistency over pairings of original and 'degraded' translations (Graham, Baldwin, Moffat, and Zobel, 2014);
- It is backwards-compatible with the style of system preference judgment used for WMT evaluations, and provides a mechanism for enabling longitudinal evaluation of MT systems;
- It is cheap: we obtained high-quality, statistically significant assessments at a cost of around US$40 per system for a given language direction and evaluation modality (that is, adequacy or fluency).

To investigate the feasibility of the approach for large-scale evaluations, we replicated the original WMT-12 human evaluation of all participating systems for two language pairs: Spanish-to-English and English-to-Spanish. Results show that our method results in high-quality data at low cost, without the use of expert assessments. Indeed, we demonstrate that it is possible to confirm the original system rankings; and also, using comparable numbers of judgments, identify a larger number of significant differences between systems.

The low cost of the method suggests that, beyond applications for cross-system evaluation in shared tasks, it may also provide a viable means for individual MT researchers to assess system improvements over a baseline. While development of the tools used by

the workers involved significant effort and refinement, this effort need not be repeated, as we make these tools available.[1] Overall, we show that carefully-gathered crowd-sourced assessments lead to more sensitive measurements than do assessments from other sources, suggesting that our methods should be used widely for the measurement of MT systems.

2 Background

2.1 Automatic Measurement of MT Systems

The development of effective mechanisms for evaluation of MT system output has long been a research objective within MT, with several of the recommendations of the early ALPAC Report (ALPAC, 1966), for example, relating to evaluation:

1. Practical methods for evaluation of translations; ...
3. Evaluation of quality and cost of various sources of translations;

In practical terms, improvements are often established through the use of an automatic metric that computes a similarity score between the candidate translation and one or more human-generated reference translations. However it is well known that automatic metrics are not necessarily a good substitute for human assessments of translation quality, and must be used with caution (Turian, Shen, and Melamed, 2003; Callison-Burch, Osborne, and Koehn, 2006; Koehn and Monz, 2006; Lopez, 2008). Particular issues include:

- There are generally many different ways of translating the same source input, and therefore comparison with a reference translation risks artificially up-scoring translations that happen to be more reference-like compared to equally-valid translations that make different lexical or structural choices; and
- Automatic metrics are generally based on lexical similarity and fail to capture meaning, either in terms of fundamental semantic infelicity (for example, the lack of a negation marker or core argument) or underlying semantic similarity but lexical divergence with a reference translation (Koehn and Monz, 2006; Lo, Addanki, Saers, and Wu, 2013).

One way of reducing bias in evaluation towards the particular decisions made in a reference translation is to source multiple reference translations and calculate an aggregated score across them (Culy and Riehemann, 2003; Madnani, Resnik, Dorr, and Schwartz, 2008). Even here, however, automatic evaluation metrics tend to focus too much on local similarity with translations, and ignore the global fidelity and coherence of the translation (Lo et al., 2013); this introduces a bias when comparing systems that are based on differing principles. Other approaches, such as HyTER (Dreyer and Marcu, 2012), aim to encode all possible correct translations in a compact reference translation network, and match the output of an MT system against this using string edit distance. Such approaches are hampered by a lack of automation in crafting the reference translations, and currently require up to two hours per sentence.

To alleviate these concerns, direct human assessments of translation quality are also collected when possible. During the evaluation of MT shared tasks, for example, human

[1] See `https://github.com/ygraham/crowd-alone`

assessments of MT outputs have been used to determine the ranking of participating systems. The same human assessments can also be used in the evaluation of automatic metrics, by comparing the degree to which automatic scores (or ranks) of translations correlate with them. This aspect of MT measurement is discussed in the following section.

2.2 Validation of Automatic Metrics

In order to validate the effectiveness of an automatic MT evaluation metric, a common approach is to measure correlation with human assessments of MT quality. A corpus of test sentences is selected, and a set of MT systems is used to translate each sentence (often, the pool of systems participating in a shared task). Human assessors are then asked to assess the quality of the translations, or a randomly selected subset of them if experimental cost is a limiting constraint. In addition, some translation assessments are repeated, to facilitate later measurement of assessment consistency levels.

Once sufficiently many sentence-level human assessments have been collected, they are used to decide a best-to-worst ranking of the participating MT systems. A range of methods for going from sentence-level human assessments to a totally-ordered system ranking have been proposed (Bojar, Ercegovcevic, and Popel, 2011; Lopez, 2012; Callison-Burch et al., 2012; Hopkins and May, 2013), with no consensus on the best aggregation method. An automatic MT evaluation metric can be used to mechanically rank the same set of systems, perhaps based on document-level scoring. The system ranking generated based on the human assessments and the automatic metric can then be compared based on correlation, using, for example, Spearman's ρ or Pearson's r, with a high correlation interpreted as evidence that the metric is sound. The robustness of this use of Spearman correlation has been questioned, as it cannot discriminate between errors with respect to varying magnitude in system scores (Graham and Baldwin, 2014; Machacek and Bojar, 2014).

Since the validity of an automatic MT evaluation measure is assessed relative to human assessments, it is vital that the human assessments are reliable. In practice, measurement of evaluation reliability is based on evaluation of intra- and inter-annotator agreement. There is a worrying trend of low agreement levels for MT shared tasks, however. For example, Cohen's κ coefficient (Cohen, 1960), is commonly used to quantify inter-annotator agreement levels and incorporates the likelihood of agreement occurring by chance, with a coefficient ranging between 0, signifying agreement at chance levels, to 1 for complete agreement. In recent WMT shared tasks, κ coefficients as low as $\kappa = 0.40$ (2011), $\kappa = 0.33$ (2012), $\kappa = 0.26$ (2013), and $\kappa = 0.37$ (2014) have been reported. Intra-annotator agreement levels have not been much better: $\kappa = 0.58$ (2011), $\kappa = 0.41$ (2012), $\kappa = 0.48$ (2013), and $\kappa = 0.52$ (2014) (Callison-Burch, Koehn, Monz, and Zaidan, 2011; Callison-Burch et al., 2012; Bojar et al., 2013, 2014). It is possible that alternative measures that capture different types of disagreement, such as Krippendorff's α, are more appropriate for the task, given that it involves multiple coders (Poesio and Artstein, 2005; Mathet et al., 2012). However, such low levels of agreement are unlikely to be wholly attributable to a deficiency in the κ coefficient as a measure of agreement, as individual annotators only weakly agree with themselves, let alone other annotators.

The lack of coherence amongst human assessments raises a critical question: *are assessments of MT evaluation metrics robust, if they are validated via low-quality human*

assessments of translation quality? One possible reaction to this question is that the automatic evaluation measures are no worse than human assessment. A more robust response is to find ways of increasing the reliability of the human assessments used as the yardstick for automatic metrics, by identifying better ways of collecting and assessing translation quality. It may be, for example, that we are not asking assessors "is this a good translation" in a form that leads to a consistent interpretation.

There has been significant effort invested in developing metrics that correlate with human assessments of translation quality. However, given the extent to which accurate human assessment of translation quality is fundamental to empirical MT, the underlying topic of finding ways of increasing the reliability of those assessments to date has received surprisingly little attention (Callison-Burch et al., 2007, 2008; Przybocki, Peterson, Bronsart, and Sanders, 2009; Callison-Burch et al., 2009, 2010; Denkowski and Lavie, 2010).

2.3 Past and Current Methodologies

The ALPAC Report (ALPAC, 1966) was one of the earliest published attempts to perform cross-system MT evaluation, in determining whether progress had been made over the preceding decade. The (somewhat anecdotal) conclusion was that:

> The reader will find it instructive to compare the samples above with the results obtained on simple, selected, text 10 years earlier ... in that the earlier samples are more readable than the later ones.

The DARPA Machine Translation Initiative of the 1990s incorporated MT evaluation as a central tenet, and periodically evaluated the three MT systems funded by the program (CANDIDE (Berger et al., 1994), PANGLOSS (Frederking et al., 1994) and LINGSTAT (Yamron, Cant, Demedts, Dietzel, and Ito, 1994)). As part of this, it examined whether post-editing of MT system output was faster than simply translating the original from scratch (White, O'Connell, and O'Mara, 1994). One of the major outcomes of the project was the proposal that *adequacy* and *fluency* be used as the primary means of human MT evaluation, supplemented by other human-assisted measurements. Adequacy is the degree to which the information in the source language string is preserved in the translation,[2] while fluency is the determination of whether the translation is a well-formed natural utterance in the target language.

Many of the large corporate machine translation systems use regression testing to establish whether new methods have a positive impact on MT quality. Annotators are asked to select which of two randomly-ordered translations they prefer, one from each system (Bond, Ogura, and Ikehara, 1995; Schwartz, Aikawa, and Quirk, 2003), often over a reference set of translation pairs (Ikehara, Shirai, and Ogura, 1994).

Approaches trialled for human judgments of translation quality include all of:

- Ordinal level scales (ranking a number of translations from best-to-worst) or direct estimates (interval-level scales) of fluency or adequacy judgments;

[2] Or, in the case of White et al. (1994), the degree to which the information in a professional translation can be found in the machine translation, as judged by monolingual speakers of the target language.

- Different lexical units (for example, sub-sentential constituents rather than whole sentences);
- Differing numbers of points on an interval-level scale (for example, having 4, 5, or 7 points);
- Displaying or not displaying interval-level scale numbering to annotators;
- Simultaneously assessing fluency and adequacy items, or separating the assessment of fluency and adequacy;
- Changing the wording of the question used to elicit judgments (for example, asking which translation is *better*, or asking which is *more adequate*);
- Including or not including the reference translation among the set being judged;
- Displaying the translation of the sentences either side of the target sentence, or not displaying any surrounding context;
- Displaying or not displaying session or overall participation meta-information to the assessor (for example, the number of translations assessed so far, the time taken so far, or the number of translations left to be assessed);
- Employing or not employing crowd-sourced judgments.

In recent years the annual WMT event has become the main forum for collection of human assessments of translation quality, despite the primary focus of the workshop being to provide a regular cross-system comparison over standardized data sets for a variety of language pairs by means of a shared translation task (Koehn and Monz, 2006; Callison-Burch et al., 2007, 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013, 2014). When WMT began in 2006, fluency and adequacy were adopted, and were also used in the LDC report (LDC, 2005) to assess participating systems in the form of direct estimates on a 5-point interval-level scale. Too few human assessments were recorded in the first year to be able to estimate the reliability of human judgments (Koehn and Monz, 2006). In 2007, the workshop measured the consistency of the human judgments, to boost the robustness of evaluation; to achieve this goal, human assessments of translations were provided by participants in the shared task.

The reliability of human judgments was then estimated by measuring levels of agreement. In addition, two new further methods of human assessment were added: (1) generate a partial ranking of translations of full sentences from best to worst (or relative preference judgments); and (2) generate a partial ranking of translations of sub-sentential source syntactic constituents from best to worst. The first of these methods has continued in recent WMT evaluations; in both cases, ties are allowed. The highest levels of agreement were reported for the sub-sentential source syntactic constituent ranking method ($\kappa_{\text{inter}} = 0.54$, $\kappa_{\text{intra}} = 0.74$), followed by the full sentence ranking method ($\kappa_{\text{inter}} = 0.37$, $\kappa_{\text{intra}} = 0.62$). The lowest agreement levels occurred for adequacy ($\kappa_{\text{inter}} = 0.31$, $\kappa_{\text{intra}} = 0.54$). Additional methods of human assessment have also been trialled, but the only method still currently used is the best-to-worst partial ranking of translations, known as "relative preference judgments".

Another finding from WMT 2007 with widespread consequences was high correlation between fluency and adequacy, which was taken as evidence for redundancy in separately assessing the two, and led to the conflation of translation quality into a single scale (Przybocki et al., 2009; Denkowski and Lavie, 2010). For example, Przybocki

et al. (2009) use, as part of their larger human evaluation, a single (7-point) scale (labeled "adequacy") to assess the quality of translations. Inter-annotator agreement for this method was $\kappa = 0.25$, even lower than the results for adequacy and fluency reported in WMT 2007 (noting that caution is required when directly comparing agreement measurements, especially over scales of varying granularity, such as 5- versus 7-point assessments).

Three recent WMT shared tasks (2011, 2012, and 2013) have limited their human assessment to partially ranking translations from best-to-worst. Even with this simpler conceptualization of the annotation task, consistency levels are still low. Agreement levels reported in, for example, WMT-12 using translation ranking were lower than those reported in 2007 with fluency and adequacy assessments.

As an additional confound, in the absence of a better alternative, WMT human evaluations are typically carried out by MT researchers participating in the shared task, rather that by independent volunteers, or by expert translators. Although the assessments of translations are blind, and the system that produced any particular translation is hidden from the judges, researchers have been shown to slightly favor translations produced by their own system (Bojar et al., 2011), highlighting the need for other methods of incorporating human evaluation into MT evaluation. Crowd-sourced human judgments eliminate that particular confound, but bring a different risk – vastly different skill and/or levels of care. Hence, until now, participating researchers have been preferred as judges to crowd-sourced assessors.

Turning to the question of what number of judgments is needed to produce statistically significant rankings, simulation experiments suggest that when assessment is in the form of relative preference judgments, this number is likely to grow quadratically in the number of participating systems (Koehn, 2012; Bojar et al., 2014). For example, up to triple the number of judgments collected in WMT-12 may have been required to identify sufficient proportions of significant differences between systems (Koehn, 2012). Although a range of models have been proposed that aim to leverage the shortcomings of relative preference judgments to identify significant differences between systems, such as Hopkins and May (2013) and Sakaguchi et al. (2014), our work instead focuses on attending to the root cause of what we believe to be the problem: the information-loss that takes place when there is no attempt to capture the degree to which one translation is better than another.

3 Simplified Monolingual Assessment

Direct (or absolute) estimates of translation quality facilitate more powerful statistical analyses of systems than do partial ordinal rankings via relative preference judgments, which, for example, cannot be readily combined into mean and median scores. Being able to combine individual direct estimate scores into a mean or median score simplifies the decision as to how an overall ranking should be concluded; moreover, the law of large numbers suggests that as increasing numbers of individual scores are collected for each system, the mean score will increasingly approach the true score. By simply increasing the number of assessments (given access to suitably skilled assessors), accuracy is improved.

3.1 Capturing Direct Estimate Assessments

To address those concerns, we propose that assessments be based on direct estimates of quality using a continuous rating scale. We further suggest that it is undesirable to assess several translations at a time, as even when performing direct estimation, there is the potential for ratings of different translations to influence each other. For example, if a translation happens to appear in the same displayed set as a high quality translation, it is likely that its score will be pushed down (Bojar et al., 2011). To minimize this kind of bias, we present a single translation per screen to human assessors, and require that each translation be considered in isolation. Assessors can only proceed to the next translation once they have committed their score for the current one, with no facility to revisit or revise earlier decisions. A third component of our proposal is to revert to the use of two-dimensional fluency and adequacy assessments, in part because doing so allows the assessments to be presented in the form of focused questions that are more likely to elicit consistent responses. In addition, by making the fluency component independent, part of the overall evaluation becomes completely reference-free, and hence not subject to possible comparison bias. We use the same continuous scale for both fluency and adequacy.

For technical reasons, the continuous rating scale is presented as a 100-point slider, with the assessor selecting a rating by moving the slider with their mouse. Strictly speaking, it is thus not continuous; but the actual rating obtained is interpreted as a real-valued number, and the proposed methodology could trivially be applied to a finer-grained scale (for example, when higher-resolution screens become commonplace). Numbering is intentionally avoided, but the slider is marked at the mid and quarterly points for rough calibration purposes. We make every effort to make the assessment task compatible with crowd-sourcing, including couching both the fluency and adequacy questions as monolingual (target language) tasks, to maximize the pool of potential assessors. We intentionally obscure the fact that the text is the product of machine translation, to eliminate that knowledge as a potential source of bias.

The sections that follow describe the assessment procedure for fluency and adequacy in detail below, and the processes employed for score standardization and filtering out poor-quality assessors. The platform we use to collect the assessments is Amazon Mechanical Turk (AMT, see `https://www.mturk.com`), in which jobs are presented to "workers" in the form of human intelligence tasks (abbreviated as "HITs"), each corresponding to a single unit that a worker accepts through the AMT interface. Each HIT is made up of either all adequacy or all fluency assessment tasks, and each task is presented on a separate screen, with no facility for workers to return to revise earlier assessments. In all cases, the translated text is presented to the worker as a bit-mapped image, in order to deter robotic workers who could, for example, screen-scrape the translated text and use some automatic method to calculate their assessment. Although the assessments in this work were collected specifically with AMT, other crowd-sourcing services could also be used or set-up for this purpose. Therefore, although precise replication of the experimental results presented in this work may rely on the existence and continuity of the AMT service, more importantly the approach itself does not.

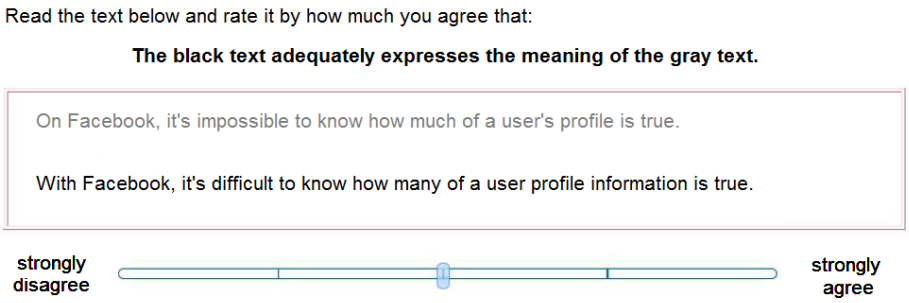

Fig. 1. Screenshot of the adequacy assessment interface, as presented to an AMT worker. All of the text is presented as an image. The slider is initially centered; workers move it to the left or right in reaction to the question. No scores or numeric information are available to the assessor.

3.2 Adequacy: Measuring Equivalence of Meaning

The first dimension of the evaluation is assessment of adequacy. Adequacy is assessed as a monolingual task, with a single translation presented per screen. The reference translation is provided in gray font at the top of the screen, with the system translation displayed below it in black. Assessors are asked to state the degree to which they agree that: *The black text adequately expresses the meaning of the gray text*.

The task is thus restructured into a less cognitively-taxing monolingual similarity-of-meaning task, under the fundamental assumption that the reference translation accurately captures the meaning of the source sentence. Once that presumption is made, it is clear that the source is not actually required for evaluation. An obvious benefit of this change is that the task now requires only monolingual speakers, without any knowledge of translation or MT. Figure 1 is a screenshot of a single adequacy assessment as posted on AMT.

Reference translations used for the purpose of adequacy assessments are those included as standard in MT test sets, and this avoids the need to, for example, attempt to generate reference translations through expert translators or crowd-sourcing. If reference translations were to be crowd-sourced, care would of course need to be taken to ensure effective quality control of this process to avoid introduction of noise into adequacy evaluations.

3.3 Reference-Free Evaluation of Fluency

The second dimension of the evaluation is assessment of fluency. Fluency assessments are presented using the translated text only, with neither the source language input text shown, nor any reference translation(s). This removes any bias towards systems that happen to produce reference-like translations, a common criticism of automatic metrics such as BLEU. It also forces our assessors to make an independent fluency judgment on the translation, with the intention of minimizing biasing from other assessments.

Figure 2 shows a screenshot of a fluency assessment, as one task within an AMT HIT of similar assessments. As with the adequacy assessments, a single translation is displayed to the human assessor at a time, which they are then asked to rate for fluency. The assessment

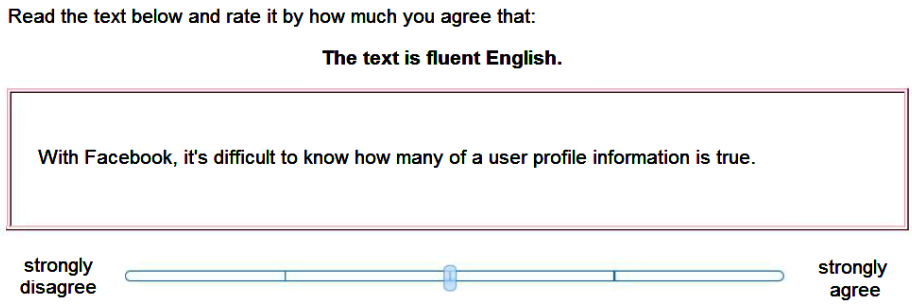

Fig. 2. Screenshot of the fluency assessment interface, as presented to an AMT worker. Many of the details are the same as for the adequacy assessment shown in Figure 1.

is carried out based on the strength of agreement with the Likert-type statement: *The text is fluent English*.

With source and reference sentences not shown, fluency assessments alone cannot be used to determine if one system is better than another. This is because it is possible for a system to produce highly-fluent text with complete disregard for the actual content of the source input. In our evaluation, fluency assessments are used as a means of breaking ties between systems that are measured to have equal adequacy, as well as a diagnostic tool that is unbiased in favor of systems that produce reference-like translations.

3.4 HIT Structure and Score Standardization

Since fluency and adequacy assessments are carried out separately, we set these assessments up as separate HITs. Both kinds of HIT contain 100 translation assessments each, one per screen. The worker is thus required to iterate through translations, rating them one at a time. Within each HIT genuine system output translations may be paired with a repeat item, a corresponding reference translation, or a "*bad_reference*" translation (a degraded version of the same translation, explained in more detail below). The aim of this selective pairing is that post-HIT completion, regardless of the overall scoring strategy of individual workers, their internal consistency at scoring better or worse translations can be examined without comparison to someone else's scores. Structuring the HIT as 100 translations allows us to manipulate the task in such a way that we have a high level of control of same-judge repeat and quality-control items, as follows.

Within a 100-translation HIT, we include: (a) 10 reference translations and 10 MT system outputs for the same source language strings as the reference translations (making a total of 20 translations); (b) 10 MT system outputs, along with a mechanically-degraded "*bad_reference*" version of each (making a total of another 20 translations; see below for details of the degradation process); and (c) another 50 MT system outputs, out of which we select 10 and repeat them verbatim (making a total of 60 translations). That is, each 100-translation HIT consists of:

- 70 MT system outputs,

- 10 reference translations, corresponding to 10 of the 70 system outputs
- 10 *bad_reference* translations, corresponding to a different 10 of the 70 system outputs,
- 10 repeat MT system outputs, drawn from the remaining 50 of the original 70 system outputs.

The role of the reference, *bad_reference*, and repeat translations is explained in the next section.

The order in which translations are assessed by a worker is controlled so that a minimum of 40 assessments intervene between each member of a pair of quality-control items (*bad_reference* versus MT system, or MT system versus MT system repeat, or reference translation versus MT system). The selection is further controlled so that each 100-translation HIT contains approximately equal numbers of randomly-selected translations from each contributing MT system, so as to provide overall balance in the number of translations that are judged for each system. That is, no matter how many HITs each worker completes, they will return roughly the same number of assessments for each of the contributing systems. This helps avoid any potential skewing of results arising from particularly harsh or lenient assessors.

To further homogenize the outputs of the different workers, the set of scores generated by each person is *standardized*. This is done in the usual way, by computing the mean and standard deviation of the scores returned by that particular assessor, and then translating each of their raw scores into a z score. The result is a set of scores that, for each assessor, has a mean of 0.0 and a standard deviation of 1.0; it is those values that are then averaged across systems to get system scores. Standardizing the scores removes any individual biases for particular sub-regions of the scale (for example, workers who tend to rate everything low or high), and also for the relative spread of scores used by a given worker (for example, workers who use the full scale versus those who use only the central region). In the results tables below we report system averages of both the raw worker responses and of the standardized values.

4 Quality Control and Assessor Consistency

The quality of a human evaluation regime can be estimated from its *consistency*: whether the same outcome is achieved if the same question is asked a second time. Two different measurements can be made: whether a judge is consistent with other assessments performed by themselves (intra-annotator agreement), and whether a judge is consistent with other judges (inter-annotator agreement).

In MT, annotator consistency is commonly measured using Cohen's κ coefficient, or some variant thereof (Artstein and Poesio, 2008). Cohen's κ is intended for use with categorical assessments, but is also commonly used with five-point adjectival-scale assessments, where the set of categories has an explicit ordering. A particular issue with five-point assessments is that there is no notion of "nearness" – a judge who assigns two neighboring intervals is awarded the same "penalty" for being different as a judge who selects two extreme values.

The κ coefficient cannot be applied to continuous data; and nor can any judge, when

given the same translation to evaluate twice on a continuous rating scale, be expected to give the same score for each assessment. A more flexible tool is thus required. Our approach to measuring assessor consistency is based on two core assumptions:

***A*:** When a consistent judge is presented with a set of assessments for translations from two systems, one of which is known to produce better translations than the other, the score sample of the better system will be significantly greater than that of the inferior system.

***B*:** When a consistent judge is presented with a set of repeat assessments, the score sample across the initial presentations will not be significantly different from the score sample across the second presentations.

We evaluate Assumption A based on the set of *bad_reference* translations and the corresponding set of MT system translations, and evaluate Assumption B based on the pairs of repeat judgments in each HIT. The idea behind *bad_reference* translations is that if words are removed from a translation to shorten it, worse adequacy judgments can be expected; and if words in a translation are duplicated, fluency ratings should suffer. The null hypothesis to be tested for each AMT worker is that the score difference for repeat judgment pairs is not less than the score difference for *bad_reference* pairs. To test statistical significance, we use the Wilcoxon rank sum test,[3] with lower p values indicating more reliable workers (that is, greater differentiation between repeat judgments and *bad_reference* pairs). We use $p < 0.05$ as a threshold of reliability, and discard all of the HITs contributed by workers who do not meet this threshold. Note that if HITs are rejected in this manner, the workers may still receive payment.

The *bad_reference* translations that are inserted into each HIT are deliberately degraded relative to their matching system translations, on the assumption that a measurable drop in the assessor's rating should be observed. For the adequacy HITs, we degrade the translation by randomly deleting a short sub-string, emulating omission of a phrase. Initial experimentation showed that for this approach to be effective we needed to delete words in rough proportion to the length of the original translation, as follows:

- for 2–3 word translations, remove 1 word;
- for 4–5 word translations, remove 2 words;
- for 6–8 word translations, remove 3 words;
- for 9–15 word translations, remove 4 words;
- for 16–20 word translations, remove 5 words;
- for translations of word length $n > 20$, remove $\lfloor n/5 \rfloor$ words.

The *bad_reference* translations for fluency HITs are created by randomly selecting two words in the translation and duplicating them elsewhere in the string, excepting adjacent to the original, string-initial and string-final positions. Preliminary experiments indicated that it was not necessary to modify the length of the randomly-selected phrase according to the length of the overall translation. Duplicating a pair of words was always sufficient to obtain the desired effect.

[3] See Graham et al. (2013) for a description of experiments using Welch's t-test, Cohen's κ and the Mann-Whitney U-test.

Automatic introduction of errors to corpora have been used for other purposes besides quality control of crowd-sourced data. Pevner and Hearst (2002) simulate errors for the purpose of analyzing evaluation metrics of thematic segmentation, while Mathet et al. (2012) use reference annotation shuffles according to different error paradigms to model the behavior of a range of agreement measures. Additionally, the automatic introduction of errors into corpora have been used to create training data for error detection applications (Bigert, 2004; Bigert, Sjöbergh, Knutsson, and Sahlgren, 2005; Brockett, Dolan, and Gamon, 2006; Dickinson, 2010; Foster, 2007; Foster and Andersen, 2009; Izumi, Uchimoto, Saiga, Supnithi, and Isahara, 2003; Lee and Seneff, 2008; Rozovskaya and Roth, 2010; Sjöbergh and Knutsson, 2005; Smith and Eisner, 2005b, 2005a; Okanohara and Tsujii, 2007; Wagner, Foster, and van Genabith, 2007, 2009; Yuan and Felice, 2013).

4.1 Crowd-Sourcing Specifics

Mechanical Turk was used directly to post HITs to crowd-sourced workers, as opposed to any intermediary service, and fluency and adequacy HITs are collected in entirely separate sessions. Since our HIT structure is somewhat unconventional, with 100 translations per HIT, it was important that information about the quantity of work involved per HIT was communicated to workers prior to their acceptance of a HIT. An additional specification was that only native speakers of the target language complete HITs, and although we do not have any way of verifying that workers adhered to this request, it is important to communicate expected language fluency levels to workers. Payment was at the rate of US$0.50 per 100-translation fluency HIT, and increased to US$0.90 per adequacy HIT – the difference because, in the latter case, HITs involved reading both a reference translation and the assessed translation. Workers were not required to complete a qualification test prior to carrying out HITs, as such qualification restrictions unfortunately do not ensure high quality work. Due to the anonymous nature of crowd-sourcing, it is of course entirely possible for workers lacking the necessary skills to employ someone else to complete his/her test. In addition, qualification tests do not provide any assurance that skilled workers don't carefully completing tests, and then aggressively optimize earnings at a later stage. In contrast, the quality control mechanism we employ does not rely on one-off tests but applies quality checks across all of the HITs provided by each worker.

Since the quality control mechanism we apply could in fact be too high a bar for some genuine workers to meet, we do not use or recommend its use as the sole basis for accepting or rejecting HITs. Indeed, some workers may simply lack the necessary literacy skills to accurately complete evaluations effectively. During collection of crowd-sourced assessments, we therefore only reject workers who we believe are almost certainly attempting to aggressively optimize earnings by gaming the system. Aggressive optimizers are identified by comparison of the worker's mean score for reference translations, genuine system outputs and bad reference translations. For example, random-clicker type aggressive optimizers commonly have extremely close mean scores for all three kinds of quality control items. An additional check is also put in place to search for score sequences within individual HITs with low variation. Another helpful tool to identify aggressive optimizers specifically with adequacy assessments, is that when a reference translation appears as the item to be assessed it is in fact identical to the reference translation displayed on screen.

	Evaluation modality	HITs		
		Total	Approved	Pass QC
ES-EN	Fluency	288	260 (90%)	152 (53%)
	Adequacy	543	410 (76%)	192 (35%)
EN-ES	Fluency	336	250 (74%)	211 (63%)
	Adequacy	278	245 (88%)	204 (73%)

Table 1. *Human Intelligence Task (HIT) approval and rejection in experiments.*

Although such items appear to be too obvious to be useful in assessments, they in fact act as a further modest hurdle for workers to meet, and caught many workers. For example, in some cases mean scores for degraded and genuine system outputs may be suspiciously close: mean scores for reference items for such workers often reveal that they rated identical items with low adequacy, and therefore can be rejected with confidence.

Not rejecting all workers whose assessments do not meet the quality control threshold unfortunately results in a set of HITs from workers whose data is not good enough to be useful in evaluations but nonetheless require payment. Table 1 shows numbers of HITs approved and rejected in experiments and corresponding numbers of HITs belonging to workers who passed quality control. Although proportions of low quality workers for whom we accept HITs varies from one language pair to the other, volumes of such workers are not so large for either language pair to be in any way considered prohibitively costly. It may be worth noting, however, that in pilot posting of HITs on Mechanical Turk, we encountered a substantial increase in numbers of workers that fall within this costly group when we attempted to increase payment levels. In particular, when payment for HITs is increased to US$1 or more per HIT, this appears to attract substantially more workers who fall into this category. Although we have otherwise complied with ethics guidelines available to researchers, such as `http://wiki.wearedynamo.org/index.php?title=Guidelines_for_Academic_Requesters`, our experience when attempting to increase payment levels is that they (unfortunately) must remain low to avoid attracting large numbers of aggressive optimizers.

Fort, Adda and Cohen (2011) identify as a main cause of low payment to workers the lack of adequate worker reputation systems: without a method of accurately targeting good workers, increased payment simply attracts larger numbers of aggressive optimizers. The quality control mechanism we propose accurately discriminates good work from bad, and in this respect, could improve the situation for workers by facilitating such a reputation system. For example, although we do not refuse payment based solely on whether or not a particular worker meets the quality control threshold we use to filter low quality data, information about the quality of work could be collected over the longer term and used to rate individual workers. This information could then be used by requesters, not as a strict cut-off, but to provide reliable fine-grained information about the quality likely to be produced by workers. Other ethical issues associated with crowd-sourcing – such as intellectual property, potential for states being deprived of tax payments, and unknown working conditions and rights of employment, due to the anonymous nature of crowd-sourcing services – remain a real concern but for the same reason are challenging to reliably investi-

gate. Gupta et al. (2014) suggest that relationship-based crowd-sourcing could potentially be more fruitful than many current modes of operation, where relationships are maintained with a group of known good workers through emailing when batches of work are ready. Such an approach is compatible with our proposed methodology.

Languages that can be evaluated effectively through crowd-sourcing are of course limited to the native languages of speakers on a particular service. In this respect, previous efforts to elicit evaluations for Czech, for example, have resulted in so few genuine responses to HITs that we believe evaluation of this language pair by the crowd alone is currently not possible (Graham, Baldwin, Harwood, Moffat, and Zobel, 2012).

5 System-Level Evaluation

We have described a method for collecting fine-grained assessments of MT quality, and for filtering out unreliable workers. To investigate the utility and robustness of the proposed methodology, we replicate the human evaluation component of the WMT-12 shared translation task for Spanish and English in both translation directions, generating fresh crowd-sourced human assessments for all participating systems. This allows the system rankings produced by the new evaluation methodology to be compared with those of the original shared task. In addition, we compute the degree to which statistically significant differences can be identified for pairs of participating systems.

5.1 Data

Samples of translations from the published WMT-12 shared task data set (Callison-Burch et al., 2012) for Spanish-to-English and English-to-Spanish translation were selected at random and evaluated by AMT workers. Twelve systems participated in the original Spanish-to-English (ES-EN) shared translation task, and eleven systems in English-to-Spanish (EN-ES). Including quality control items, we collected a total of approximately 62k human fluency judgments and 82k human adequacy judgments.

Table 2 shows the number and percentage of workers who passed quality control, and the percentage of workers with no significant difference between mean scores for exact repeat items. Consistent with previous findings (Graham et al., 2013), the quality of English-speaking workers on AMT appears to be lower than for Spanish-speaking workers. Note that while we use the statistical tests described in the previous section to determine which HITs we use for the system evaluation, we do not use them as the basis for accepting/rejecting HITs – that is, for determining whether a given worker should be paid for performing a HIT. Our method of quality control is a high bar to reach, and workers might act in good faith and yet still not be consistent enough to meet the significance threshold we imposed. Instead, we individually examined mean score differences for reference translation, system output, and *bad_reference* pairs, and declined payment only when there was no doubt that the response was either automatic or extremely careless.

After quality-control filtering over *bad_reference* pairs based on Assumption A, approximately 36k fluency judgments and 41k adequacy judgments remained. When we subsequently applied quality control based on Assumption B using exact-repeat translations, no additional HITs were filtered for adequacy for either language direction (that is, there were

	Evaluation modality	Workers			Translations	
		Total	A holds	A&B hold	Total	A&B hold
ES-EN	Fluency	102	54 (53%)	53 (52%)	29k	15k (52%)
	Adequacy	319	94 (30%)	94 (30%)	54k	19k (35%)
EN-ES	Fluency	37	22 (60%)	21 (57%)	33k	21k (64%)
	Adequacy	45	21 (46%)	21 (46%)	28k	20k (71%)

Table 2. *Numbers of workers and translations, before and after quality control (broken down based on Assumption A only, and both Assumptions A&B).*

	Adequacy	Fluency	WMT-12
Spanish-to-English	25.5	12.3	50.9
English-to-Spanish	29.0	12.0	–

Table 3. *Average time per assessment (seconds) with fluency and adequacy direct estimate assessments and* WMT-12 *relative preference assessments.*

no instances of significant differences between score distributions for exact repeat items), whereas for fluency, two additional workers, one for each language direction, were filtered out.

5.2 Evaluation of Spanish-to-English Systems

Table 4 shows mean raw and standardized human adequacy and fluency scores for each participating WMT-12 system. Rows in the table are ordered from best to worst according to mean standardized adequacy scores, with mean standardized fluency scores used as a secondary key where adequacy scores agree to two decimal places. Systems are anonymized with ordered names according to their rank order derived from subsequent significance tests (Figure 3), with "I" unused to avoid confusion with the number "1". Out-of-sequence system names in this table indicate a divergence between final conclusions based on combined adequacy/fluency significance tests (Figure 3) and the numeric ordering based on mean scores alone (this table).

Table 3 shows the average per-translation time taken by assessors who passed our quality control requirements, and for the assessors involved in WMT-12.[4] A comparison reveals a reduction in time taken to assess translations by approximately half when direct estimates are elicited by our monolingual set-up compared to relative preference judgments.

For this language pair the raw and standardized mean scores are closely correlated, a by-product of the structure of the HITs, which ensures that there is a relatively even distribution of systems across workers, as described in Section 3.4. If it had not been possible to put this constraint on assessments – for example, if assessments from two separate groups of systems and human judges were being combined – score standardization is likely to

[4] No information about WMT-12 times for English-to-Spanish is currently available.

	Adequacy			Fluency		
	z	raw	n	z	raw	n
System A	0.21	65.7	1,328	0.32	63.8	1,002
System C	0.15	62.0	1,287	0.17	59.2	995
System B	0.12	61.7	1,270	0.24	61.4	975
System D	0.12	61.7	1,274	0.14	58.5	1,041
System E	0.10	60.6	1,256	0.16	59.2	1,006
System F	0.03	59.4	1,264	0.00	54.0	1,075
System H	0.03	59.4	1,232	−0.05	53.0	954
System G	0.02	59.2	1,272	−0.03	52.6	1,006
System J	−0.03	58.3	1,344	0.04	56.1	1,035
System K	−0.05	57.1	1,273	−0.06	51.6	999
System L	−0.17	53.5	1,288	−0.21	47.3	1,039
System M	−0.54	42.6	1,272	−0.70	33.1	1,033

Table 4. *Spanish-to-English mean human adequacy and fluency scores ("z" is the mean standardized z-score, and "n" is the total number of judgments for that system after quality filtering is applied).*

have a greater effect. In general, significance tests on standardized scores are more robust than on raw scores, and if a genuine difference in behavior exists, it is more likely to be identified in the standardized data.

Significance tests are used to estimate the degree to which rankings between pairs of systems are likely to have occurred simply by chance. For our human-sourced data, we apply a one-sided Wilcoxon rank-sum test to standardized score distributions for pairs of systems. Results of significance tests for all pairs of systems for Spanish-to-English for standardized adequacy and fluency assessments are shown as the first and second heat maps in the lower half of Figure 3.

The third heat map in the lower half of the figure is a combined adequacy and fluency test, constructed as follows: if system X's adequacy score is significantly greater than that of system Y at some p value, then the combined conclusion is that X is significantly better than Y at that p-value. If the outcome of a significance test for a pair of systems' adequacy score distributions is not significant at the desired significance level, we consider this to be a "tie" in terms of adequacy, in which case the conclusions of a significance test on fluency assessments are used to derive the system ordering. In that case conclusions from the significance test for fluency scores for that pair of systems should be taken as the overall outcome. For example, adequacy tests for Systems B and C show no significant difference in Figure 3, but tests on additional fluency assessments reveal that System B produces translations that are judged to be significantly more fluent than those of System C. System B is therefore regarded as being superior to System C. Note that this strategy of tie-breaking is only applied when there is no significant difference in adequacy at $p < 0.05$. As has already been noted, in our test environment fluency alone cannot be used as criteria to rank systems. For example, although System J achieves higher fluency than System G in

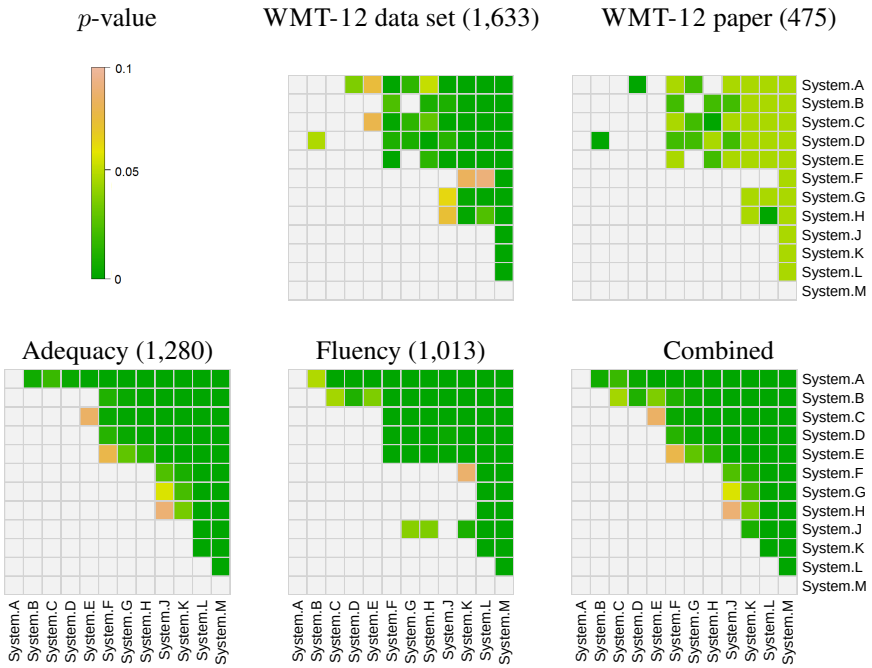

Fig. 3. Spanish-to-English significance test outcomes for each method of human evaluation. Colored cells indicate that the scores of the row i system are significantly greater than those of the column j system. The average number of judgments per system is shown in parentheses. The top row shows the official results from WMT-12; the bottom row show the results based on our method.

Figure 3, System G achieves significantly higher adequacy than System J, and in this pair it is System G that is regarded as superior.

The upper part of Figure 3 shows significance test results for the WMT-12 evaluation, computed on the published data set ("WMT-12 data set"), and on the official results ("WMT-12 paper"), the latter generated from the former by filtering via agreement with expert items (Callison-Burch et al., 2012). Since the original WMT-12 human evaluation took the form of relative preference judgments, we are unable to use the Wilcoxon rank sum test, and instead use a sign test to test for statistical significance in both the WMT-12 published data set and the official results. Since the rank order of systems is known at the time of significance testing, the application of a single instance of a one-tailed test to each pair of competing systems is appropriate, testing if the score distribution of the higher ranking system is significantly greater than that of the lower ranking system. All p-values reported in Figure 3 are for one-tailed tests, and therefore the heat map matrices in Figure 3 have maximally $n(n-1)/2$ filled cells.

Comparing the combined adequacy-fluency significance matrix with those of the original relative preference evaluation, it is clear that a similar system ordering has emerged; but our methodology yields a higher proportion of significant differences between systems, and provides a more conclusive system ranking with fewer uncertainties. As a single ex-

	Adequacy			Fluency		
	z	raw	n	z	raw	n
System A	0.18	65.7	1,532	0.16	59.9	1,463
System B	0.05	61.8	1,486	0.13	58.5	1,541
System C	0.05	60.4	1,465	0.12	58.5	1,579
System E	0.05	61.4	1,499	0.07	57.2	1,565
System F	0.03	61.5	1,491	0.04	55.8	1,588
System D	0.02	60.3	1,498	0.09	57.4	1,501
System J	−0.03	58.3	1,482	−0.13	49.5	1,531
System K	−0.07	56.7	1,414	−0.14	49.9	1,545
System G	−0.08	57.0	1,517	−0.08	51.6	1,553
System H	−0.10	57.3	1,469	−0.08	51.5	1,507
System L	−0.11	57.2	1,467	−0.17	49.5	1,507

Table 5. *English-to-Spanish mean human adequacy and fluency scores ("z" is the mean standardized z-score, and "n" is the total number of judgments for that system after quality filtering is applied).*

ception to the overall pattern, System D was found to significantly outperform System B in the WMT evaluation, while we found no significant difference in adequacy, and in contrast to WMT results, that System B in fact outperformed System D in terms of fluency. The new evaluation data allows System A to be declared the outright winner for this language pair, with significantly higher adequacy scores than all other participating systems. This level of separation was not possible using the data generated by the original relative preference-based evaluation.

5.3 Evaluation of English-to-Spanish Systems

Table 5 shows mean scores of each participating system in the WMT-12 English-to-Spanish task, based on our evaluation methodology, with the table rows ordered using the mean standardized adequacy scores as a primary key, and the fluency scores as a secondary key. The same system ranking and naming convention is used as for Table 4.

For this language pair, differences in system rankings with respect to raw and standardized scores are again not large. Nonetheless, the standardized score ranking pushes System C up two places, as its raw mean of 60.4 is below those of System E (61.4) and System F (61.5). When computed on the (less-biased) standardized scores the mean score for System C is now (roughly) equal to System E (0.05) and higher than System F (0.03).

As before, significance testing provides insight into the relative performance of the set of systems. Figure 4 parallels Figure 3, and includes results from the original WMT-12 shared task human evaluation used to determine the official results ("WMT-12 paper"), and results based on the published WMT-12 data set ("WMT-12 data set"); to provide a meaningful comparison, all tests in Figure 4 are one-tailed. In this translation environment, the fluency scores allow several ties in adequacy to be broken, including the five-way tie for second place between Systems B, C, D, E, and F. For English-to-Spanish, the new

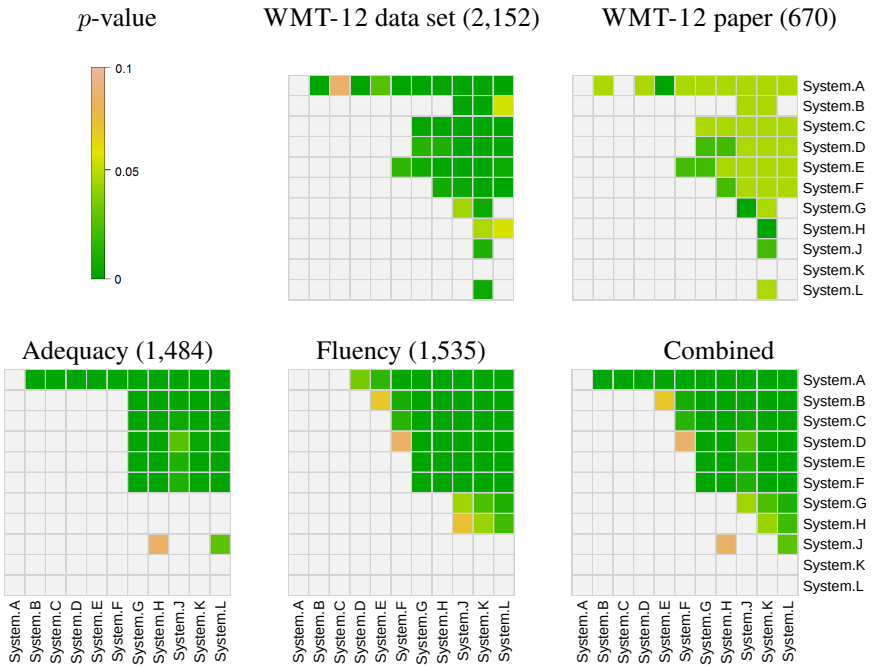

Fig. 4. English-to-Spanish significance test outcomes for each method of human evaluation. Colored cells indicate that the scores of the row i system are significantly greater than those of the column j system. The average number of judgments per system is shown in parentheses. The top row shows the official results from WMT-12; the bottom row show the results based on our method.

evaluation methodology again produces almost a superset of official WMT-12 results but with more certainty, this time with three exceptions. In the original results, System E was shown to beat System F, System J to beat System K, and System L to beat System K. None of these conclusions are supported by the results of the new methodology.

Overall, System A is identified as an outright winner in our experimentation. In contrast, the original relative preference-based human evaluation was unable to identify any single system that significantly outperformed all others. Moreover, the outright winner identified for this English-to-Spanish language pair, System A, is the same system identified for Spanish-to-English. In addition, the systems taking second and third place for the two translation directions are also matched. That is, System B and System C each represent the same system for both language pairs, and take second and third place respectively. With these three exceptions, the anonymized names of systems for Spanish-to-English and English-to-Spanish do not correspond.

5.4 Varying the Number of Assessments

The results described above suggest that our crowd-sourced approach to gathering judgments achieves higher levels of system discrimination than the approach used at WMT-12.

p	Spanish-to-English		English-to-Spanish	
	Relative preference	Direct estimate	Relative preference	Direct estimate
0.05	19.7%	69.7%	16.4%	65.5%
0.01	6.1%	56.1%	5.5%	52.7%
0.001	–	39.4%	–	41.8%

Table 6. *Proportions of significant differences between system pairs identified at different significance thresholds, using the* WMT-12 *relative preference judgments, and the new direct estimate method.*

In addition, there was a relatively high level of overall concord between the fluency and adequacy judgments, allowing us to break ties in adequacy based on fluency. However, our experiments made use of more judgments than the WMT-12 evaluation, and it could be that the larger pool of judgments collected from the AMT workers is the factor responsible for the greater certainty. We investigate that possibility by reducing the number of HITs used in our analyses.

Since our method diverges considerably from the original WMT-12 approach to collecting assessments, determining the correct number of judgments to use for the purpose of comparison is not entirely straightforward. In the WMT-12 evaluation, five translations are assessed per screen, and ordered from best to worst to generate ten pairwise labels. To order those five competing translations, it seems likely that each translation is read more than once. If the average number of readings is two, then ten labels are generated each time ten translations are read. On the other hand, in the direct estimate evaluation, ten labels are produced by reading and assessing ten translations, with only limited re-reading anticipated. We therefore compare methods based on numbers of labels produced by each method, as a reasonable estimate of the effort involved in generating them.

Table 6 shows the proportion of significant differences identified by our direct estimate-based evaluation compared to those of the original WMT-12 relative preference evaluation, for equal numbers of judgments. The number of judgments for the direct estimate method is limited to the number in the official WMT-12 data set, namely, 475 judgments per system for the 12 Spanish-to-English systems, and 670 judgments per system for the 11 English-to-Spanish systems. In the latter case, the subset of direct estimate judgments was made by taking HITs in the order they were completed by AMT workers.

These results show that our method reveals substantially higher proportions of significant differences between pairs of participating systems compared to the original relative preference evaluation. At $p < 0.05$, for example, for the same number of judgments, our direct estimate evaluation produces approximately 3.5 times the number of significant differences between pairs of Spanish-to-English systems, and close to 4 times as many significant differences for English-to-Spanish systems. At the higher confidence level of $p < 0.01$, the relative proportion of significant differences detected by the two methods diverges even further, with our direct estimate method identifying 9–10 times the proportion of significant differences identified by relative preference judgments.

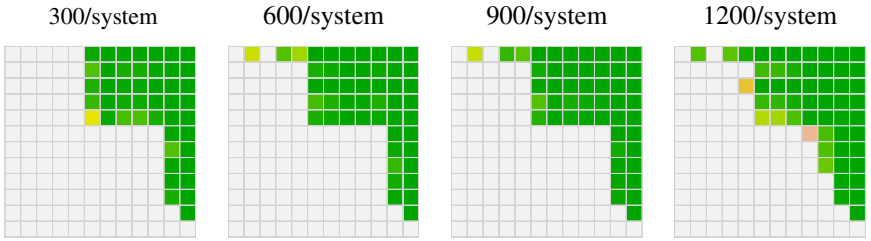

Fig. 5. Standardized adequacy significant differences between pairs of systems for increasing numbers of judgments per system, sampled according to earliest HIT submission time, for the twelve Spanish-to-English WMT-12 systems. These four heat maps can be directly compared to the lower-left heat map in Figure 3, which is constructed using an average of 1280 judgments per system.

The heat maps in Figure 5 provide additional insight into the relationship between the number of judgments per system and the level of statistical significance for different system pairings, based on direct estimation methodology. As few as 300 crowd-generated adequacy judgments per system is sufficient for mid-range and low-range systems to be separated from the better-performing ones. Note that even the last 80 judgments collected, from 1200 per system on average to 1280 per system on average (which is what is shown in Figure 3), are helpful in distinguishing between systems.

The aim of the evaluation we have detailed is to provide an accurate mechanism for evaluation of MT on the system-level. Although ratings for translations assessed on a continuous rating scale can be expected to contain random variations in individual scores, when ratings are collected for a sufficiently large sample of translations belonging to a given system, positive and negative random errors present in individual assessments cancel out to produce accurate mean and median scores for systems. For evaluations to be accurate at the segment-level, a different approach is required, where assessments are repeated per segment as opposed to per system; see Graham et al. (2015) for further details.

6 Conclusion

We have presented a new methodology for human evaluation of machine translation quality. To our knowledge, this is the first method that relies entirely on assessments sourced from the crowd, in our case using Amazon's Mechanical Turk. Our approach is based on actively removing sources of bias, including mechanisms to accommodate assessors with consistent individual scoring strategies. By restructuring the task as an assessment of monolingual similarity of meaning, assessing individual translations, and separating fluency and adequacy, the task was made substantially less cognitively taxing, and allowed participation by much larger pools of workers.

To assess the feasibility of our methodology for large-scale MT evaluation, we replicated the WMT-12 shared task human evaluation for two language pairs, using the system translations released by the task organizers. Our results show that capturing direct estimates of translation quality on a continuous rating scale leads to more informative judgments that reflect not only the fact that one translation is superior to another, but also the degree to

which it was preferred. In addition, direct estimates provide the advantage of a straightforward combination of individual standardized scores for translations into mean system scores that, given large numbers of assessments, are more precise. The same logic does not apply to the same degree to methods based on combinations of relative preference judgments. Once the quality control mechanisms were applied, high consistency between workers was observed. Score standardization at the worker level was also helpful.

With our new assessments, score distributions for translations of competing systems provide useful discrimination between systems, and we identified substantially higher proportions of significant differences. More conclusive rankings not only provide greater insight into the relative performance of MT systems, but also establish a better foundation for evaluation of automatic metrics. The evaluation methodology is scalable, and highly efficient and cost-effective: the judgments used in this paper were collected at approximately US$40 per system, for each language pair and evaluation modality (English-to-Spanish or Spanish-to-English, and adequacy or fluency). With greater consistency per assessment than previous approaches, and clearer rankings across the systems considered, we have answered in the affirmative the question we posed as the title of this paper: machine translation systems can indeed be evaluated by the crowd alone.

Acknowledgements

This work was supported by the Australian Research Council's *Discovery Projects* Scheme (grant DP110101934) and Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (`www.adaptcentre.ie`) at Trinity College Dublin.

References

ALPAC. 1966. *Languages and machines: Computers in translation and linguistics.* A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council.

Artstein, R., and Poesio, M. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–96.

Berger, A., Brown, P., Della Pietra, S., Della Pietra, V., Gillett, J., Lafferty, J., Mercer, R. , Printz, H. and Ureš, L. 1994. The Candide system for machine translation. In *Proceedings of the 1994 Human Language Technology Workshop*, pp. 157–62. Plainsboro, NJ.

Bigert, J. 2004. Probabilistic detection of context-sensitive spelling errors. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pp. 1633–6. Lisbon, Portugal.

Bigert, J., Sjöbergh, J., Knutsson, O., and Sahlgren, M. 2005. Unsupervised evaluation of parser robustness. In *Proceedings of the Sixth International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 142–54. Mexico City, Mexico: Springer.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R. and Specia, L. 2013. Findings of the 2013 Workshop on Sta-

tistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pp. 1–44. Sofia, Bulgaria.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L. and Tamchyna, A. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58. Baltimore, MA.

Bojar, O., Ercegovcevic, M., Popel, M., and Zaidan, O. 2011. A grain of salt for the WMT manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 1–11. Edinburgh, Scotland.

Bond, F., Ogura, K., and Ikehara, S. 1995. Possessive pronouns as determiners in Japanese-to-English machine translation. In *Proceedings of the Second Paclic Association for Computational Linguistics Conference: PACLING-95*, pp. 32–8. Brisbane, Australia.

Brockett, C., Dolan, W., and Gamon, M. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics*, pp. 249–56. Sydney, Australia.

Callison-Burch, C. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 286–95. Suntec, Singapore.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 136–58. Prague, Czech Republic.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pp. 70–106. Columbus, OH.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and Zaidan, O. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, pp. 17–53. Uppsala, Sweden.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pp. 10–51. Quebec, Canada.

Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pp. 1–28. Athens, Greece.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 22–64. Edinburgh, Scotland.

Callison-Burch, C., Osborne, M., and Koehn, P. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the Eleventh European Chapter of the Association for Computational Linguistics*, pp. 249–56. Trento, Italy.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Culy, C., and Riehemann, S. Z. 2003. The limits of n-gram translation evaluation metrics. In *Proceedings of the Ninth Machine Translation Summit*, pp. 71–8. New Orleans, LA.

Denkowski, M., and Lavie, A. 2010. Choosing the right evaluation for machine translation: An examination of annotator and automatic metric performance on human judgement tasks. In *Proceedings of the Twelfth Conference of the Association of Machine Translation in the Americas.* Denver, CO.

Dickinson, M. 2010. Generating learner-like morphological errors in Russian. In *Proceedings of the Twenty-Third International Conference on Computational Linguistics*, pp. 259–67. Beijing, China.

Dreyer, M., and Marcu, D. 2012. Hyter: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 162–71. Quebec, Canada.

Fort, K., Adda, G., and Cohen, K. B. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, *37*(2), 413–20.

Foster, J. 2007. Treebanks gone bad. *International Journal of Document Analysis and Recognition (IJDAR)*, *10*(3-4), 129–45.

Foster, J., and Andersen, Ø. 2009. GenERRate: generating errors for use in grammatical error detection. In *Proceedings of the Fourth Workshop on Innovative Use of Natural Language Processing for Building Educational Applications*, pp. 82–90. Suntec, Singapore.

Frederking, R., Nirenburg, S., Farwell, D., Helmreich, S., Hovy, E., Knight, K., Beale, S., Domashnev, C., Attardo, D., Grannes, D. and Brown, R. 1994. Integrating translations from multiple sources within the Pangloss Mark III machine translation system. In *Proceedings of the First Conference of the Association of Machine Translation in the Americas*, pp. 73–80. Columbia, MA.

Graham, Y., and Baldwin, T. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 172–6. Doha, Qatar.

Graham, Y., Baldwin, T., Harwood, A., Moffat, A., and Zobel, J. 2012. Measurement of progress in machine translation. In *Proceedings of the Australasian Language Technology Workshop*, pp. 70–8. Dunedin, New Zealand: Australasian Language Technology Association.

Graham, Y., Baldwin, T., and Mathur, N. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies*, pp. 1183–91. Denver, CO.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings Seventh Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 33–41. Sofia, Bulgaria.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. 2014. Is machine translation getting better over time? In *Proceedings of the Fourteenth European Chapter of the Association for Computational Linguistics*, pp. 443–51. Gothenburg, Sweden.

Graham, Y., Mathur, N., and Baldwin, T. 2014. Randomized significance tests in machine

translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 266–74. Baltimore, MA.

Gupta, N., Martin, D., Hanrahan, B., and O'Neill, J. 2014. Turk-life in India. In *Proceedings of the Eighteenth International Conference on Supporting Group Work*, pp. 1–11. Sanibel Island, FL.

Hopkins, M., and May, J. 2013. Models of translation competitions. In *Proceedings of the Fifty-First Annual Meeting of the Association for Computational Linguistics*, pp. 1416–24. Sofia, Bulgaria.

Ikehara, S., Shirai, S., and Ogura, K. 1994. Criteria for evaluating the linguistic quality of Japanese to English machine translations. *Journal of the Japanese Society for Artificial Intelligence*, *9*. (In Japanese)

Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T., and Isahara, H. 2003. Automatic error detection in the Japanese learners' English spoken data. In *Proceedings of the Forty-First Annual Meeting of the Association for Computational Linguistics*, pp. 145–8. Sapporo, Japan.

Koehn, P. 2012. Simulating human judgment in machine translation evaluation campaigns. In *Proceedings of the Ninth International Workshop on Spoken Language Translation*, pp. 179–84. Miami, Florida.

Koehn, P., and Monz, C. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings of the First Workshop on Statistical Machine Translation*, pp. 102–21. New York City, NY.

Kumar, S., and Byrne, W. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting*, pp. 169–76. Boston, MA.

LDC. 2005. *Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations* (Tech. Rep.). Linguistic Data Consortium. (Revision 1.5)

Lee, J., and Seneff, S. 2008. Correcting misuse of verb forms. In *Proceedings of the Forty-Sixth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 174–82. Columbus, OH.

Lo, C., Addanki, K., Saers, M., and Wu, D. 2013. Improving machine translation by training against an automatic semantic frame based evaluation metric. In *Proceedings of the Fifty-First Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 375–81. Sofia, Bulgaria.

Lopez, A. 2008. Statistical machine translation. *ACM Computing Survey*, *40*(3), 1–49.

Lopez, A. 2012. Putting human machine translation systems in order. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pp. 1–9. Quebec, Canada.

Machacek, M., and Bojar, O. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 293–301. Baltimore, MA.

Madnani, N., Resnik, P., Dorr, B. J., and Schwartz, R. 2008. Are multiple reference translations necessary? investigating the value of paraphrased reference translations in parameter optimization. In *Proceedings of the Tenth Conference of the Association of Machine Translation in the Americas.* Waikiki, HI.

Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S. and Zweigenbaum, P. 2012. Manual corpus annotation: Giving meaning to the evaluation metrics. In *Proceedings of the Twenty-Fourth International Conference on Computational Linguistics: Posters*, pp. 809–18. Mumbai, India.

Och, F. J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the Forty-First Annual Meeting of the Association for Computational Linguistics*, pp. 160–67. Sapporo, Japan.

Okanohara, D., and Tsujii, J. 2007. A discriminative language model with pseudo-negative samples. In *Proceedings of the Forty-Fifth Annual Meeting of the Association for Computational Linguistics*, pp. 73–80. Prague, Czech Republic.

Pevzner, L., and Hearst, M. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, *28*(1), 19–36.

Poesio, M., and Artstein, R. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of Frontiers in Corpus Annotations II: Pie in the Sky*, pp. 76–83. Ann Arbor, MI.

Przybocki, M., Peterson, K., Bronsart, S., and Sanders, G. 2009. The NIST 2008 metrics for machine translation challenge: Overview, methodology, metrics and results. *Machine Translation*, *23*(2-3), 71–103.

Rozovskaya, A., and Roth, D. 2010. Training paradigms for correcting errors in grammar and usage. In *Proceedings of Human Language Technologies: The Eleventh Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 154–62. Los Angeles, CA.

Sakaguchi, K., Post, M., and Durme, B. V. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Fifty-Second Annual Meeting of the Association for Computational Linguistics*, pp. 1–11. Baltimore, MA.

Schwartz, L., Aikawa, T., and Quirk, C. 2003. Disambiguation of English PP attachment using multilingual aligned data. In *Proceedings of the Ninth Machine Translation Summit.* New Orleans, LA.

Sjöbergh, J., and Knutsson, O. 2005. Faking errors to avoid making errors: Very weakly supervised learning for error detection in writing. In *Proceedings of Recent Advances in Natural Language Processing.* Borovets, Bulgaria.

Smith, N., and Eisner, J. 2005a. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the Forty-Third Annual Meeting of the Association for Computational Linguistics*, pp. 354–62. Ann Arbor, MI.

Smith, N., and Eisner, J. 2005b. Guiding unsupervised grammar induction using contrastive estimation. In *Proceedings of the International Joint Conference on Artificial Intelligence Workshop on Grammatical Inference Applications*, pp. 73–82. Edinburgh, Scotland.

Turian, J. P., Shen, L., and Melamed, I. D. 2003. Evaluation of Machine Translation and its Evaluation. In *Proceedings of the Ninth Machine Translation Summit*, pp. 386–93. New Orleans, LA.

Wagner, J., Foster, J., and van Genabith, J. 2007. A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. In *Proceedings of the Joint Meeting of the Conference on Empirical Methods in Natu-*

ral Language Processing and the Conference on Computational Natural Language Learning, pp. 112–21. Prague, Czech Republic.

Wagner, J., Foster, J., and van Genabith, J. 2009. Judging grammaticality: Experiments in sentence classification. *Computer-Assisted Language Instruction Consortium Journal*, *26*(3), 474–90.

White, J. S., O'Connell, T., and O'Mara, F. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association of Machine Translation in the Americas*, pp. 193–205. Columbia, MA.

Yamron, J., Cant, J., Demedts, A., Dietzel, T., and Ito, Y. 1994. The automatic component of the LINGSTAT machine-aided translation system. In *Proceedings of the 1994 Human Language Technology Workshop*, pp. 163–8. Plainsboro, NJ.

Yuan, Z., and Felice, M. 2013. Constrained grammatical error correction using statistical machine translation. In *Proceedings of the Seventh Conference on Computational Natural Language Learning*, pp. 52–61. Sofia, Bulgaria.