Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk

Chris Callison-Burch
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, Maryland
ccb@cs.jhu.edu

Abstract

Manual evaluation of translation quality is generally thought to be excessively time consuming and expensive. We explore a fast and inexpensive way of doing it using Amazon's Mechanical Turk to pay small sums to a large number of non-expert annotators. For $10 we redundantly recreate judgments from a WMT08 translation task. We find that when combined non-expert judgments have a high-level of agreement with the existing gold-standard judgments of machine translation quality, and correlate more strongly with expert judgments than Bleu does. We go on to show that Mechanical Turk can be used to calculate human-mediated translation edit rate (HTER), to conduct reading comprehension experiments with machine translation, and to create high quality reference translations.

1 Introduction

Conventional wisdom holds that manual evaluation of machine translation is too time-consuming and expensive to conduct. Instead, researchers routinely use automatic metrics like Bleu (Papineni et al., 2002) as the sole evidence of improvement to translation quality. Automatic metrics have been criticized for a variety of reasons (Babych and Hartley, 2004; Callison-Burch et al., 2006; Chiang et al., 2008), and it is clear that they only loosely approximate human judgments. Therefore, having people evaluate translation output would be preferable, if it were more practical.

In this paper we demonstrate that the manual evaluation of translation quality is not as expensive or as time consuming as generally thought. We use Amazon's Mechanical Turk, an online labor market that is designed to pay people small sums of money to complete *human intelligence tests* – tasks that are difficult for computers but easy for people. We show that:

- Non-expert annotators produce judgments that are very similar to experts and that have a stronger correlation than Bleu.
- Mechanical Turk can be used for complex tasks like human-mediated translation edit rate (HTER) and creating multiple reference translations.
- Evaluating translation quality through reading comprehension, which is rarely done, can be easily accomplished through creative use of Mechanical Turk.

2 Related work

Snow et al. (2008) examined the accuracy of labels created using Mechanical Turk for a variety of natural language processing tasks. These tasks included word sense disambiguation, word similarity, textual entailment, and temporal ordering of events, but not machine translation. Snow et al. measured the quality of non-expert annotations by comparing them against labels that had been previously created by expert annotators. They report inter-annotator agreement between expert and non-expert annotators, and show that the average of many non-experts converges on performance of a single expert for many of their tasks.

Although it is not common for manual evaluation results to be reported in conference papers, several large-scale manual evaluations of machine translation quality take place annually. These include public forums like the NIST MT Evaluation Workshop, IWSLT and WMT, as well as the project-specific Go/No Go evaluations for the DARPA GALE program. Various types of human judgments are used. NIST collects 5-point fluency and adequacy scores (LDC, 2005), IWSLT and

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 286–295,
Singapore, 6-7 August 2009.

WMT collect relative rankings (Callison-Burch et al., 2008; Paul, 2006), and DARPA evaluates using HTER (Snover et al., 2006). The details of these are provided later in the paper. Public evaluation campaigns provide a ready source of gold-standard data that non-expert annotations can be compared to.

3 Mechanical Turk

Amazon describes its Mechanical Turk web service[1] as *artificial* artificial intelligence. The name and tag line refer to a historical hoax from the 18th century where an automaton appeared to be able to beat human opponents at chess using a clockwork mechanism, but was, in fact, controlled by a person hiding inside the machine. The Mechanical Turk web site provides a way to pay people small amounts of money to perform tasks that are simple for humans but difficult for computers. Examples of these Human Intelligence Tasks (or HITs) range from labeling images to moderating blog comments to providing feedback on relevance of results for a search query.

Anyone with an Amazon account can either submit HITs or work on HITs that were submitted by others. Workers are sometimes referred to as "Turkers" and people designing the HITs are "Requesters." Requesters can specify the amount that they will pay for each item that is completed. Payments are frequently as low as $0.01. Turkers are free to select whichever HITs interest them.

Amazon provides three mechanisms to help ensure quality: First, Requesters can have each HIT be completed by multiple Turkers, which allows higher quality labels to be selected, for instance, by taking the majority label. Second, the Requester can require that all workers meet a particular set of qualications, such as sufficient accuracy on a small test set or a minimum percentage of previously accepted submissions. Finally, the Requester has the option of rejecting the work of individual workers, in which case they are not paid.

The level of good-faith participation by Turkers is surprisingly high, given the generally small nature of the payment.[2] For complex undertakings like creating data for NLP tasks, Turkers do not have a specialized background in the subject, so there is an obvious tradeoff between hiring individuals from this non-expert labor pool and seeking out annotators who have a particular expertise.

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

[2] For an analysis of the demographics of Turkers and why they participate, see: http://behind-the-enemy-lines.blogspot.com/2008/03/mechanical-turk-demographics.html

4 Experts versus non-experts

We use Mechanical Turk as an inexpensive way of evaluating machine translation. In this section, we measure the level of agreement between expert and non-expert judgments of translation quality. To do so, we recreate an existing set of gold-standard judgments of machine translation quality taken from the Workshop on Statistical Machine Translation (WMT), which conducts an annual large-scale human evaluation of machine translation quality. The experts who produced the gold-standard judgments are computational linguists who develop machine translation systems.

We recreated all judgments from the WMT08 German-English News translation task. The output of the 11 different machine translation systems that participated in this task was scored by ranking translated sentences relative to each other. To collect judgements, we reproduced the WMT08 web interface in Mechanical Turk and provided these instructions:

Evaluate machine translation quality *Rank each translation from Best to Worst relative to the other choices (ties are allowed). If you do not know the source language then you can read the reference translation, which was created by a professional human translator.*

The web interface displaced 5 different machine translations of the same source sentence, and had radio buttons to rate them.

Turkers were paid a grand total of $9.75 to complete nearly 1,000 HITs. These HITs exactly replicated the 200 screens worth of expert judgments that were collected for the WMT08 German-English News translation task, with each screen being completed by five different Turkers. The Turkers were shown a source sentence, a reference translation, and translations from five MT systems. They were asked to rank the translations relative to each other, assigning scores from best to worst and allowing ties.

We evaluate non-expert Turker judges by measuring their inter-annotator agreement with the WMT08 expert judges, and by comparing the correlation coefficient across the rankings of the machine translation systems produced by the two sets of judges.

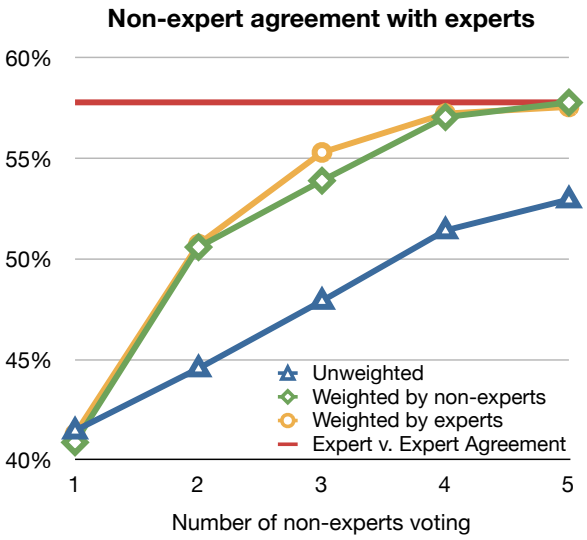

Figure 1: Agreement on ranking translated sentences increases as more non-experts vote. Weighting non-experts' votes based on agreement with either experts or other non-expert increases it up further. Five weighted non-experts reach the top line agreement between experts.

Combining ranked judgments Each item is redundantly judged by five non-experts. We would like to combine of their judgments into a single judgment. Combining ranked judgments it is more complicated than taking simple majority vote. We use techniques from *preference voting*, in which voters rank a group of candidates in order of preference. To create an ordering from the the ranks assigned to the systems by multiple Turkers, we use Schulze's method (Schulze, 2003). It is guaranteed to correctly pick the winner that is preferred pairwise over the other candidates. It further allows a complete ranking of candidates to be constructed, making it a suitable method for combining ranked judgments.

Figure 1 shows the effect of combining non-experts judgments on their agreement with experts. Agreement is measured by examining each pair of translated sentence and counting when two annotators both indicated that A $>$ B, A $<$ B, or A $=$ B. Chance agreement is $\frac{1}{3}$. The top line indicates the inter-annotator agreement between WMT08 expert annotators, who agreed with each other 58% of the time. When we have only a single non-expert annotator's judgment for each item, the agreement with experts is only 41%. As we increase the number of non-experts to five, their agreement with experts improves to 53%, if their votes are counted equally.

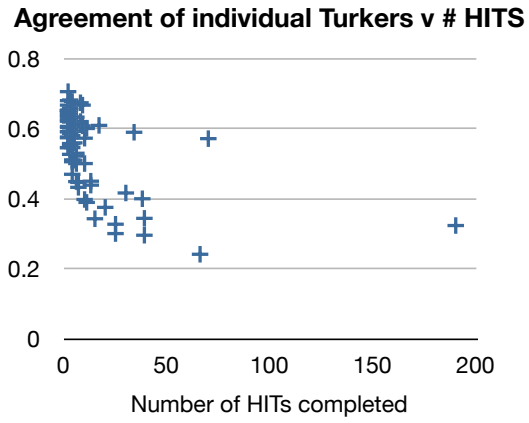

Figure 2: The agreement of individual Turkers with the experts. The most prolific Turker performed barely above chance, indicating random clicking. This suggests that users who contribute more tend to have lower quality.

Weighting votes Not all Turkers are created equal. The quality of their works varies. Figure 2 shows the agreement of individual Turkers with expert annotators, plotted against the number of HITs they completed. The figure shows that their agreement varies considerably, and that Turker who completed the most judgments was among the worst performing.

To avoid letting careless annotators drag down results, we experimented with weighted voting. We weighted votes in two ways:

- Votes were weighted by measuring agreement with experts on the 10 initial judgments made. This would be equivalent to giving Turkers a pretest on gold standard data and then calibrating their contribution based on how well they performed.
- Votes were weighted based on how often one Turker agreed with the rest of the Turkers over the whole data set. This does not require any gold standard calibration data. It goes beyond simple voting, because it looks at a Turker's performance over the entire set, rather than on an item-by-item basis.

Figure 1 shows that these weighting mechanisms perform similarly well. For this task, deriving weights from agreement with other non-experts is as effective as deriving weights from experts. Moreover, by weighting the votes of five Turkers,

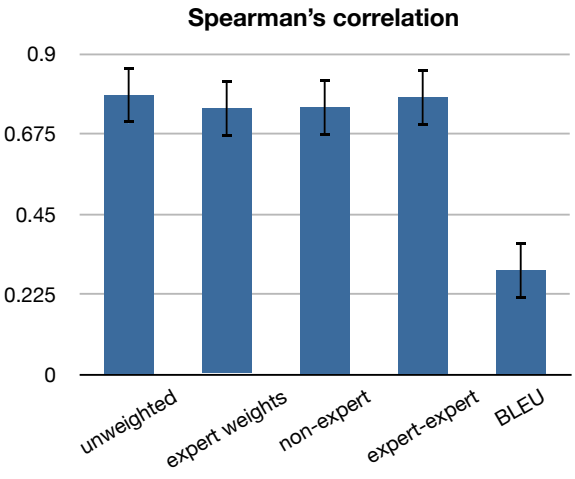

Figure 3: Correlation with experts' ranking of systems. All of the different ways of combining the non-expert judgments perform at the upper bound of expert-expert correlation. All correlate more strongly than Bleu.

we are able to achieve the same rate of agreement with experts as they achieve with each other.

Correlation when ranking systems In addition to measuring agreement with experts at the sentence-level, we also compare non-expert *system-level* rankings with experts. Following Callison-Burch et al. (2008), we assigned a score to each of the 11 MT systems based on how often its translations were judged to be better than or equal to any other system. These scores were used to rank systems and we measured Spearman's ρ against the system-level ranking produced by experts.

Figure 3 shows how well the non-expert rankings correlate with expert rankings. An upper bound is indicated by the *expert-expert* bar. This was created using a five-fold cross validation where we used 20% of the expert judgments to rank the systems and measured the correlation against the rankings produced by the other 80% of the judgments. This gave a ρ of 0.78. All ways of combining the non-expert judgments resulted in nearly identical correlation, and all produced correlation within the range of with what we would experts to.

The rankings produced using Mechanical Turk had a much stronger correlation with the WMT08 expert rankings than the Blue score did. It should be noted that the WMT08 data set does not have multiple reference translations. If multiple references were used that Bleu would likely have stronger correlation. However, it is clear that the cost of hiring professional translators to create multiple references for the 2000 sentence test set would be much greater than the $10 cost of collecting manual judgments on Mechanical Turk.

5 Feasibility of more complex evaluations

In this section we report on a number of creative uses of Mechanical Turk to do more sophisticated tasks. We give evidence that Turkers can create high quality translations for some languages, which would make creating multiple reference translations for Bleu less costly than using professional translators. We report on experiments evaluating translation quality with HTER and with reading comprehension tests.

5.1 Creating multiple reference translations

In addition to evaluating machine translation quality, we also investigated the possibility of using Mechanical Turk to create additional reference translations for use with automatic metrics like Bleu. Before trying this, we were skeptical that Turkers would have sufficient language skills to produce translations. Our translation HIT had the following instructions:

Translate these sentences *Your task is to translate 10 sentences into English. Please make sure that your English translation:*

- *Is faithful to the original in both meaning and style*
- *Is grammatical, fluent, and natural-sounding English*
- *Does not add or delete information from the original text*
- *Does not contain any spelling errors*

When creating your translation, please:

- *Do not use any machine translation systems*
- *You may look up a word on wordreference.com if you do not know its translation*

Afterwards, we'll ask you a few quick questions about your language abilities.

We solicited translations for 50 sentences in French, German, Spanish, Chinese and Urdu, and designed the HIT so that five Turkers would translate each sentence.

Filtering machine translation Upon inspecting the Turker's translations it became clear that many had ignored the instructions, and had simply cut-and-paste machine translation rather then translating the text themselves. We therefore set up a second HIT to filter these out. After receiving the

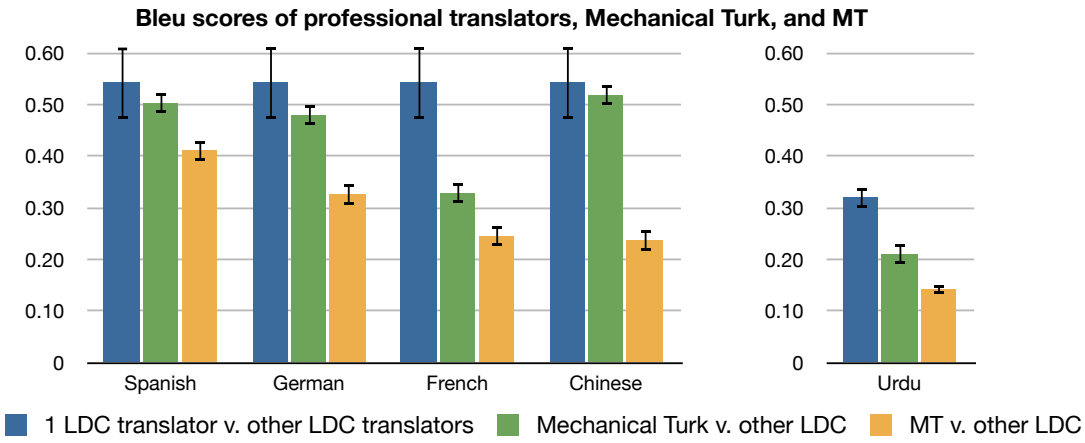

Figure 4: Bleu scores quantifying the quality of Turkers' translations. The chart shows the average Bleu score when one LDC translator is compared against the other 10 translators (or the other 2 translators in the case of Urdu). This gives an upper bound on the expected quality. The Turkers' translation quality falls within a standard deviation of LDC translators for Spanish, German and Chinese. For all languages, Turkers produce significantly better translations than an online machine translation system.

translations, we had a second group of Turkers clean the results.

Detect machine translation *Please use two online machine translation systems to translate the text into English, and then copy-and-paste the translations into the boxes below. Finally, look at a list of translations below and click on the ones that look like they came from the online translation services.*

We automatically excluded Turkers whose translations were flagged 30% of the time or more.

Quality of Turkers' translations Our 50 sentence test sets were selected so that we could compare the translations created by Turkers to translations commissioned by the Linguistics Data Consortium. For the Chinese, French, Spanish, and German translations we used the the Multiple-Translation Chinese Corpus.[3] This corpus has 11 reference human translations for each Chinese source sentence. We had bilingual graduate students translate the first 50 English sentences of that corpus into French, German and Spanish, so that we could re-use the multiple English reference translations. The Urdu sentences were taken from the NIST MT Eval 2008 Urdu-English Test Set[4] which includes three distinct English translations for every Urdu source sentence.

Figure 4 shows the Turker's translation quality in terms of the Bleu metric. To establish an upper bound on expected quality, we determined what the Bleu score would be for a professional translator when measured against other professionals. We calculated a Bleu score for each of the 11 LDC translators using the other 10 translators as the reference set. The average Bleu score for LDC2002T01 was 0.54, with a standard deviation of 0.07. The average Bleu for the Urdu test set is lower because it has fewer reference translations.

To measure the Turkers' translation quality, we randomly selected translations of each sentence from Turkers who passed the Detect MT HIT, and compared them against the same sets of 10 reference translations that the LDC translators were compared against. We randomly sampled the Turkers 10 times, and calculated averages and standard deviations for each source language. Figure 4 the Bleu scores for the Turkers' translations of Spanish, German and Chinese are within the range of the LDC translators. For all languages, the quality is significantly higher than an online machine translation system. We used Yahoo's Babelfish for Spanish, German, French and Chinese,[5] was likely and Babylon for Urdu.

Demographics We collected demographic information about the Turkers who completed the translation task. We asked how long they had spoken the source language, how long they had spo-

[3]LDC catalog number LDC2002T01

[4]LDC catalog number LDC2009E11

[5]We also compared against Google Translate, but excluded the results since its average Bleu score was better than the LDC translators, likely because the test data was used to train Google's statistical system.

Spanish		
Native lang	English (7 people), Spanish (2), English-Spanish bilingual, Portuguese	English, Hindi
Country	USA (7 people), Mexico (3), Brazil,	USA (2)
Spanish level	30+ years (2 people), 15 years (2), 6 years, 2 years (2), whole life (4)	18 years, 4 years
English level	15 years (3), whole life (9)	whole life , 15 years
German		
Native lang	German (3), Turkish (2), Italian, Danish, English, Norwegian, Hindi	Marathi, Tamil, Hindi, English
Country	Germany (3), USA, Italy, China, Denmark, Turkey, Norway, India	USA (2), India (2)
German level	20 years (2), 10 years (3), 5 years (2), 2 years, whole life (3)	10 years, 1 year (2)
English level	20+ years (4), 10-20 years (5) whole life	whole life (2), 15-20 years (2)
French		
Native lang	English (9 people), Portuguese, Hindi	English (2)
Country	USA (6), Israel, Singapore, UK, Brazil, India	USA (2)
French level	20+ years (4 people), 8-12 years (4), 5 years (2), 2 years	10 years, 1 years, 6 years
English level	whole life (9), 20 years, 15 years	whole life (2),
Chinese		
Native lang	Hindi (2)	English (3) Hindi, Marathi, Tamil
Country	India (2)	India (3), USA (3)
Chinese level	2 years, 1 year	3 years, 2 years, none
English level	18 years, 20+ years	16 years, whole life (2)
Urdu		
Native lang	Urdu (6 people)	Tamil (2), Hindi, Telugu
Country	Pakistan (3), Bahrain, India, Saudi Arabia	India (4)
Urdu level	whole life (6 people)	2 years, 1 year, never (2)
English level	20+ years (5), 15 years (2), 10 years	10+ years (5), 5 years

Table 1: Self-reported demographic information from Turkers who completed the translation HIT. The statistics on the left are for people who appeared to do the task honestly. The statistics on the right are for people who appeared to be using MT (marked as using it 20% or more in the Detect MT HIT).

ken English, what their native language was, and where they lived. Table 1 gives their replies.

Cost and speed We paid Turkers $0.10 to translate each sentence, and $0.006 to detect whether a sentence was machine translated. The cost is low enough that we could create a multiple reference set quite cheaply; it would cost less than $1,000 to create 4 reference translations for 2000 sentences.

The time it took for the 250 translations to be completed for each language varied. It took less than 4 hours for Spanish, 20 hours for French, 22.5 hours for German, 2 days for Chinese, and nearly 4 days for Urdu.

5.2 HTER

Human-mediated translation edit rate (HTER) is the official evaluation metric of the DARPA GALE program. The evaluation is conducted annually by the Linguistics Data Consortium, and it is used to determine whether the teams participating the program have met that year's benchmarks. These evaluations are used as a "Go / No Go" determinant of whether teams will continue to receive funding. Thus, each team have a strong incentive to get as good a result as possible under the metric.

Each of the three GALE teams encompasses multiple sites and each has a collection of machine translation systems. A general strategy employed by all teams is to perform system combination over these systems to produce a synthetic translation that is better than the sum of its parts (Matusov et al., 2006; Rosti et al., 2007). The contribution of each component system is weighted by the expectation that it will produce good output. To our knowledge, none of the teams perform their own HTER evaluations in order to set these weights.

We evaluated the feasibility of using Mechanical Turk to perform HTER. We simplified the official GALE post-editing guidelines (NIST and LDC, 2007). We provided these instructions:

Edit Machine Translation *Your task is to edit the machine translation making as few changes as possible so that it matches the meaning of the human translation and is good English. Please follow these guidelines:*

- *Change the machine translation so that it has the same meaning as the human translation.*
- *Make the machine translation into intelligible English.*
- **Use as few edits as possible.**
- *Do not insert or delete punctuation simply to follow traditional rules about what is "proper."*
- *Please* **do not** *copy-and-paste the human translation into the machine translation.*

	Number of editors					
System	0	1	2	3	4	5
google.fr-en	.44	.29	.24	.22	.20	.19
google.de-en	.48	.34	.30	.28	.25	.24
rbmt5.de-en	.53	.41	.33	.28	.27	.25
geneva.de-en	.65	.56	.50	.48	.45	.45
tromble.de-en	.77	.75	.74	.73	.71	.70

Table 2: HTER scores for five MT systems. The edit rate decreases as the number of editors increases from zero (where HTER is simply the TER score between the MT output and the reference translation) and five.

We displayed 10 sentences from a news article. In one column was the reference English translation, in the other column were text boxes containing the MT output to be edited. To minimize the edit rate, we collected edits from five different Turkers for every machine translated segment. We verified these with a second HIT were we prompted Turkers to:

Judge edited translations *First, read the reference human translation. After that judge the edited machine translation using two criteria:*

- *Does the edited translation have the same meaning as the reference human translation?*
- *Is it acceptable English? Some small errors are OK, so long as its still understandable.*

For the final score, we choose the edited segment which passed the criteria and which minimized the edit distance to the unedited machine translation output. If none of the five edits was deemed to be acceptable, then we used the edit distance between the MT and the reference.

Setup We evaluated five machine translation systems using HTER. These systems were selected from WMT09 (Callison-Burch et al., 2009). We wanted a spread in quality, so we took the top two and bottom two systems from the German-English task, and the top system from the French-English task (which significantly outperformed everything else). Based on the results of the WMT09 evaluation we would expect the see the following ranking from the least edits to the most edits: google.fr-en, google.de-en, rbmt5.de-en, geneva.de-en and tromble.de-en.

Results Table 2 gives the HTER scores for the five systems. Their ranking is as predicted, indicating that the editing is working as expected. The table reports averaged scores when the five annotators are subsampled. This gives a sense of how much each additional editor is able to minimize the score for each system. The difference between the TER score with zero editors, and the HTER five editors is greatest for the rmbt5 system, which has a delta of .29 and is smallest for jhu-tromble with .07.

5.3 Reading comprehension

One interesting technique for evaluating machine translation quality is through reading comprehension questions about automatically translated text. The quality of machine translation systems can be quantified based on how many questions are answered correctly.

Jones et al. (2005) evaluated translation quality using a reading comprehension test the Defense Language Proficiency Test (DLPT), which is administered to military translators. The DLPT contains a collection of foreign articles of varying levels of difficulties, and a set of short answer questions. Jones et al used the Arabic DLPT to do a study of machine translation quality, by automatically translating the Arabic documents into English and seeing how many human subjects could successfully pass the exam.

The advantage of this type of evaluation is that the results have a natural interpretation. They indicate how understandable the output of a machine translation system is better than Bleu does, and better than other manual evaluation like the relative ranking. Despite this advantage, evaluating MT through reading comprehension hasn't caught on, due to the difficulty of administering it and due to the fact that the DLPT or similar tests are not publicly available.

We conducted a reading comprehension evaluation using Mechanical Turk. Instead of simply administering the test on Mechanical Turk, we used it for all aspects from test creation to answer grading. Our procedure was as follows:

Test creation We posted human translations of foreign news articles, and ask Tukers to write three questions and provide sample answers. We gave simple instructions on what qualifies as a good reading comprehension question.

Reading comprehension test *Please read the short newspaper article, and then write three reading comprehension questions about it, giving sample answers for each of your questions. Good reading comprehension questions:*

- *Ask about why something happened or why someone did something.*
- *Ask about relationships between people or things.*
- *Should be answerable in a few words.*

Poor reading comprehension questions:

- *Ask about numbers or dates.*
- *Only require a yes/no answer.*

Question selection We posted the questions for each article back to Mechanical Turk, and asked other Turkers to vote on whether each question was a good and to indicate if it was redundant with any other questions in the set. We sorted questions to maximize the votes and minimized redundancies using a simple perl script, which discarded questions below a threshold, and eliminated all redundancies.

Taking the test We posted machine translated versions of the foreign articles along with the questions, and had Turkers answer them. We ensured that no one would see multiple translations of the same article.

Answer questions about a machine translated text *You will answer questions about an article that has been automatically translated from another language into English. The translation contains many errors, but the goal is to see how understandable it is. Please do your best to guess at the right answers to the questions. Please:*

- *Read through the automatically translated article.*
- *Answer the questions listed below, using just a few words.*
- *Give your best guess at the answers, even if the translation is hard to understand.*
- *Don't use any other information to answer the questions.*

Grading the answers We aggregated the answers and used Mechanical Turk to grade them. We showed the human translation of the article, one question, the sample answer, and displayed all answers to it. After the Turkers graded the answers, we calculated the percentage of questions that were answered correctly for each system.

Turkers created 90 questions for 10 articles, which were subsequently filtered down to 47 good questions, ranging from 3–6 questions per article. 25 Turkers answered questions about each translated article. To avoid them answering the questions multiple times, we randomly selected which system's translation was shown to them. Each system's translation was displayed an average of 5 times per article. As a control, we had three Turkers answer the reading comprehension questions using the reference translation.

System	% Correct Answers
reference	0.94
google.fr-en	0.85
google.de-en	0.80
rbmt5.de-en	0.77
geneva.de-en	0.63
jhu-tromble.de-en	0.50

Table 3: The results of evaluating the MT output using a reading comprehension test

Table 3 gives the percent of questions that were correctly answered using each of the different systems' outputs and using the reference translation. The ranking is exactly what we would expect, based on the HTER scores and on the human evaluation of the systems in WMT09. This again helps to validate that the reading comprehension methodology. The scores are more interpretable than Blue scores and than the WMT09 relative rankings, since it gives an indication of how understandable the MT output is.

Appendix A shows some sample questions and answers for an article.

6 Conclusions

Mechanical Turk is an inexpensive way of gathering human judgments and annotations for a wide variety of tasks. In this paper we demonstrate that it is feasible to perform manual evaluations of machine translation quality using the web service. The low cost of the non-expert labor found on Mechanical Turk is cheap enough to collect redundant annotations, which can be utilized to ensure translation quality. By combining the judgments of many non-experts we are able to achieve the equivalent quality of experts.

The suggests that manual evaluation of translation quality could be straightforwardly done to validate performance improvements reported in conference papers, or even for mundane tasks like tracking incremental system updates. This challenges the conventional wisdom which has long held that automatic metrics must be used since manual evaluation is too costly and time-consuming.

We have shown that Mechanical Turk can be used creatively to produce quite interesting things.

We showed how a reading comprehension test could be created, administered, and graded, with only very minimal intervention.

We believe that it is feasible to use Mechanical Turk for a wide variety of other machine translated tasks like creating word alignments for sentence pairs, verifying the accuracy of document- and sentence-alignments, performing non-simulated active learning experiments for statistical machine translation, even collecting training data for low resource languages like Urdu.

The cost of using Mechanical Turk is low enough that we might consider attempting quixotic things like human-in-the-loop minimum error rate training (Zaidan and Callison-Burch, 2009), or doubling the amount of training data available for Urdu.

Acknowledgments

This research was supported by the EuroMatrix-Plus project funded by the European Commission, and by the US National Science Foundation under grant IIS-0713448. The views and findings are the author's alone.

A Example reading comprehension questions

Actress Heather Locklear arrested for driving under the influence of drugs

The actress Heather Locklear, Amanda on the popular series Melrose Place, was arrested this weekend in Santa Barbara (California) after driving under the influence of drugs. A witness saw her performing inappropriate maneuvers while trying to take her car out of a parking space in Montecito, as revealed to People magazine by a spokesman for the Californian Highway Police. The witness stated that around 4.30pm Ms. Locklear "hit the accelerator very roughly, making excessive noise and trying to take the car out from the parking space with abrupt back and forth maneuvers. While reversing, she passed several times in front of his sunglasses." Shortly after, the witness, who at first, apparently had not recognized the actress, saw Ms. Locklear stopping in a nearby street and leaving the vehicle.

It was this person who alerted the emergency services, because "he was concerned about Ms. Locklear's life." When the patrol arrived, the police found the actress sitting inside her car, which was partially blocking the road. "She seemed confused," so the policemen took her to a specialized centre for drugs and alcohol and submitted her a test. According to a spokesman for the police, the actress was cooperative and excessive alcohol was ruled out from the beginning, even if "as the officers initially observed, we believe Ms. Locklear was under the influences drugs." Ms. Locklear was arrested under suspicion of driving under the influence of some - unspecified substance, and imprisoned in the local jail at 7.00pm, to be released some hours later. Two months ago, Ms. Locklear was released from a specialist clinic in Arizona where she was treated after an episode of anxiety and depression.

4 questions were selected

- Why did the bystander call emergency services?
  He was concerned for Ms. Locklear's life.
- Why was Heather Locklear arrested in Santa Barbara?
  Because she was driving under the influence of drugs
- Where did the witness see her acting abnormally?
  Pulling out of parking in Montecito
- Where was Ms. Locklear two months ago?
  She was at a specialist clinic in Arizona.

5 questions were excluded as being redundant

- What was Heather Locklear arrested for?
  Driving under the influence of drugs
- Where was she taken for testing?
  A specialized centre for drugs and alcohol
- Why was Heather Locklear arrested?
  She was arested on suspicion of driving under the influence of drugs.
- Why did the policemen lead her to a specialized centre for drugs and alcohol
  Because she seemed confused.
- For what was she cured for two months ago?
  She was cured for anxiety and depression.

Answers to *Where was Ms. Locklear two months ago?* that were judged to be correct:

Arizona hospital for treatment of depression; at a treatmend clinic in Arizona; in the Arizona clinic being treated for nervous breakdown; a clinic in Arizona; Arizona, under treatment for depression; She was a patient in a clinic in Arizona undergoing treatment for anxiety and depression; In an Arizona mental health facility ; A clinic in Arizona.; In a clinic being treated for anxiety and depression.; at an Arizona clinic

These answers were judged to be incorrect: *Locklear was retired in Arizona; Arizona; Arizona; in Arizona; Ms.Locklaer were laid off after a treatment out of the clinic in Arizona.*

References

Bogdan Babych and Anthony Hartley. 2004. Extending the Bleu MT evaluation method with frequency weightings. In *Proceedings of ACL*.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *Proceedings of EACL*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT08)*.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT09)*, March.

David Chiang, Steve DeNeefe, Yee Seng Chan, and Hwee Tou Ng. 2008. Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proceedings of EMNLP*.

Douglas Jones, Wade Shen, Neil Granoien, Martha Herzog, and Clifford Weinstein. 2005. Measuring translation quality by testing English speakers with a new defense language proficiency test for Arabic. In *Proceedings of the 2005 International Conference on Intelligence Analysis*.

LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5.

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation for multiple machine translation systems using enhanced hypothesis alignment. In *Proceedings of EACL*.

NIST and LDC. 2007. Post editing guidelines for GALE machine translation evaluation. Guidelines developed by the National Institute of Standards and Technology (NIST), and the Linguistic Data Consortium (LDC).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Michael Paul. 2006. Overview of the IWSLT 2006 evaluation campaign. In *Proceedings of International Workshop on Spoken Language Translation*.

Antti-Veikko Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie Dorr. 2007. Combining outputs from multiple machine translation systems. In *Proceedings of HLT/NAACL*.

Markus Schulze. 2003. A new monotonic and clone-independent single-winner election method. *Voting Matters*, (17), October.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*.

Omar F. Zaidan and Chris Callison-Burch. 2009. Feasibility of human-in-the-loop minimum error rate training. In *Proceedings of EMNLP*.