Is Machine Translation Getting Better over Time?

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

Abstract

Recent human evaluation of machine translation has focused on relative preference judgments of translation quality, making it difficult to track longitudinal improvements over time. We carry out a large-scale crowd-sourcing experiment to estimate the degree to which state-of-the-art performance in machine translation has increased over the past five years. To facilitate longitudinal evaluation, we move away from relative preference judgments and instead ask human judges to provide direct estimates of the quality of individual translations in isolation from alternate outputs. For seven European language pairs, our evaluation estimates an average 10-point improvement to state-of-the-art machine translation between 2007 and 2012, with Czech-to-English translation standing out as the language pair achieving most substantial gains. Our method of human evaluation offers an economically feasible and robust means of performing ongoing longitudinal evaluation of machine translation.

1 Introduction

Human evaluation forms the foundation on which empirical machine translation (MT) is based, whether human judges are employed directly to evaluate system output, or via the use of automatic metrics – themselves only validated through correlation with human judgments. Achieving consistent human evaluation is not easy, however. Annual evaluation campaigns conduct large-scale human assessment but report ever-decreasing levels of judge consistency – when given the same pair of translations to repeat-assess even expert human judges will worryingly often contradict both the preference judgment of other judges and even their own earlier preference (Bojar et al., 2013). For this reason, human evaluation has been targeted within the community as an area in need of attention, with increased efforts to develop more reliable methodologies.

One standard platform for human evaluation is WMT shared tasks and assessments have (since 2007) taken the form of ranking five alternate system outputs in order from best to worst (Bojar et al., 2013). This method has been shown to produce more consistent judgments compared to fluency and adequacy judgments on a five-point scale (Callison-Burch et al., 2007). However, relative preference judgments have been criticized for being a simplification of the real differences between translations, not sufficiently taking into account the large number of different types of errors of varying severity that occur in translations (Birch et al., 2013). Relative preference judgments do not take into account the degree to which one translation is better than another – there is no way of knowing if a winning system produces far better translations than all other systems, or if that system would have ranked lower if the severity of its inferior translation outputs were taken into account.

Rather than directly aiming to increase human judge consistency, some methods instead increase the number of reference translations available to automatic metrics. HTER (Snover et al., 2006) employs humans to post-edit each system output, creating individual human-targeted reference translations which are then used as the basis for computing the translation error rate. HyTER, on the other hand, is a tool that facilitates creation of very large numbers of reference translations (Dreyer and Marcu, 2012). Although both approaches increase fairness compared to automatic metrics that use a single generic reference translation, even human post-editors will inevitably vary

in the way they post-edit translations, and the process of creating even a single new reference translation for each system output is often too resource-intensive to be used in practice.

With each method of human evaluation, a trade-off exists between annotation time and the number of judgments collected. At one end of the spectrum, the WMT human evaluation collects large numbers of quick judgments (approximately 3.5 minutes per screen, or 20 seconds per label) (Bojar et al., 2013).[1] In contrast, HMEANT (Lo and Wu, 2011) uses a more time-consuming fine-grained semantic-role labeling analysis at a rate of approximately 10 sentences per hour (Birch et al., 2013). But even with this detailed evaluation methodology in place, human judges are inconsistent (Birch et al., 2013).

Although the trend appears to be toward more fine-grained human evaluation of MT output, it remains to be shown that this approach leads to more reliable system rankings – with a main reason to doubt this being that far fewer judgments will inevitably be possible. We take a counter-approach and aim to maintain the speed by which assessments are collected in shared task evaluations, but modify the evaluation set-up in two main ways: (1) we structure the judgments as monolingual tasks, reducing the cognitive load involved in assessing translation quality; and (2) we apply judge-intrinsic quality control and score standardization, to minimize noise introduced when crowd-sourcing is used to leverage numbers of assessments and to allow for the fact that human judges will vary in the way they assess translations. Assessors are regarded as reliable as long as they demonstrate consistent judgments across a range of different quality translations.

We elicit direct estimates of quality from judges, as a quantitative estimate of the magnitude of each attribute of interest (Steiner and Norman, 1989). Since we no longer look for relative preference judgments, we revert back to the original fluency and adequacy criteria last used in WMT 2007 shared task evaluation. Instead of five-point fluency/adequacy scales, however, we use a (100-point) continuous rating scale, as this facilitates more sophisticated statistical analyses of score distributions for judges, including worker-intrinsic quality control for crowd-sourcing. The latter does not depend on agreement with experts, and is made possible by the reduction in information-loss when a continuous scale is used. In addition, translations are assessed in isolation from alternate system outputs, so that judgments collected are no longer relative to a set of five translations. This has the added advantage of eliminating the criticism made of WMT evaluations that systems sometimes gain advantage from luck-of-the-draw comparison with low quality output, and vice-versa (Bojar et al., 2011).

Based on our proposed evaluation methodology, human judges are able to work quickly, on average spending 18 and 13 seconds per single segment adequacy and fluency judgment, respectively. Additionally, when sufficiently large volumes of such judgments are collected, mean scores stabilize and reveal significant differences for individual systems. Furthermore, since human evaluation takes the form of direct estimates instead of relative preference judgments, our evaluation introduces the possibility of large-scale longitudinal human evaluation. We demonstrate the value of longitudinal evaluation by investigating the improvement made to state-of-the-art MT over a five year time period (between 2007 and 2012) using best participating WMT shared task system output. Since it is likely that the test data used for shared tasks has varied in difficulty over this time period, we additionally propose a simple mechanism for scaling system scores relative to task difficulty.

Using the proposed methodology for measuring longitudinal change in MT system performance, we conclude that, for the seven European language pairs we evaluate, MT has made an average 10% improvement over the past 5 years. Our method uses non-expert monolingual judges via a crowd-sourcing portal, with fast turnaround and at relatively modest cost.

2 Monolingual Human Evaluation

There are several reasons why the assessment of MT quality is difficult. Ideally, each judge should be a native speaker of the target language, while at the same time being highly competent in the source language. Genuinely bilingual people are rare, however. As a result, judges are often people with demonstrated skills in the target language, and a working knowledge – often self-assessed – of the source language. Adding to the complexity is the discipline that is required: the task is cog-

[1] WMT 2013 reports 361 hours of labor to collect 61,695 labels with approximately one screen of five pairwise comparisons each yielding a set of 10 labels.

nitively difficult and time-consuming when done properly. The judge is, in essence, being asked to decide if the supplied translations are what they would have generated if they were asked to do the same translation.

The assessment task itself is typically structured as follows: the source segment (a sentence or a phrase), plus five alternative translations and a "reference" translation are displayed. The judge is then asked to assign a rank order to the five translations, from best to worst. A set of pairwise preferences are then inferred, and used to generate system rankings, without any explicit formation of stand-alone system "scores".

This structure introduces the risk that judges will only compare translations against the reference translation. Certainly, judges will vary in the degree to which they rely on the reference translation, making this a factor contributing to inter-judge inconsistency. For instance, even when experts judges undertake assessments, how often will a judge use the reference translation as a substitute for reading the source input or not even read the source input at all? And if crowd-sourcing is used can we really expect high proportions of Turkers to put the additional effort into reading and understanding the source input when a reference translation (probably in their native language) is displayed. If human assessors occasionally use the source input and at other times do not, or considerably vary the degree to which they use the source input string, this is likely to contribute to inconsistent judgments. We therefore trial assessments of adequacy in which the source input is not displayed to human judges. We structure assessments as a monolingual task and pose them in such a way that the focus is on comparing the *meaning* of reference translations and system outputs.[2]

We therefore ask human judges to assess the degree to which the system output conveys the same meaning as the reference translation. In this way, we focus the human judge indirectly on the question we wish to answer when assessing MT: *does the translation convey the meaning of the source?* The fundamental assumption of this approach is that the reference translation accurately captures the meaning of the source; once that presumption is made, it is clear that the source is not required during the evaluation.

Benefits of this change are that the task is both easier to describe to novice judges, and easier to answer, and that it requires only monolingual speakers opening up the evaluation to a vastly larger pool of genuinely qualified workers.

With this set-up in place for adequacy, we also re-introduce a fluency assessment. Fluency ratings can be carried out without the presence of a reference, reducing any remnant bias towards reference translations in the evaluation setup. That is, we propose a judgment regime in which each task is presented as a two-item fluency and adequacy judgment, evaluated separately, and with adequacy restructured into a monolingual "similarity of meaning" task.

When fluency and adequacy were originally used for human evaluation each rating used a 5-point adjective scale (Callison-Burch et al., 2007). However, adjectival scale labels are problematic and ratings have been shown to be highly dependent on exact wording of descriptors (Seymour et al., 1985). Alexandrov (2010) provides a summary of the extensive problems associated with the use of adjectival scale labels, including bias resulting from positively- and negatively-worded items not being true opposites of one another, and items intended to have neutral intensity in fact proving to have unique conceptual meanings.

It is often the case, however, that the question could be restructured so that the rating scale no longer requires adjectival labels, by posing the question as a statement such as *The text is fluent English* and asking the human assessor to specify how strongly they agree or disagree with that statement. The scale and labels can then be held constant across experimental set-ups for all attributes evaluated – meaning that if the scale is still biased in some way it will be equally so across all set-ups.

3 Assessor Consistency

One way of estimating the quality of a human evaluation regime is to measure its *consistency*: whether or not the same outcome is achieved if the same question is asked a second time. In MT, annotator consistency is commonly measured using Cohen's kappa coefficient, or some variant thereof (Artstein and Poesio, 2008). Originally developed as a means of establishing assessor independence, it is now commonly used in the reverse sense, with high numeric values being used as ev-

[2] This dimension of the assessment is similar but not identical to the monolingual adequacy assessment in early NIST evaluation campaigns (NIST, 2002).

idence of agreement. Two different measurements can be made – whether a judge is consistent with other judgments performed by themselves (intra-annotator agreement), and whether a judge is consistent with other judges (inter-annotator agreement).

Cohen's kappa is intended for use with categorical judgments, but is also commonly used with five-point adjectival-scale judgments, where the set of categories have an explicit ordering between them. One particular issue with five-point assessments is that score standardization cannot be applied. As such, a judge who assigns two neighboring intervals is awarded the same "penalty" for being "different" as the judge who chooses the extremities. The kappa coefficient cannot be directly applied to many-valued interval or continuous data.

This raises the question of how we should evaluate assessor consistency when a continuous rating scale is in place. No judge, when given the same translation to judge twice on a continuous rating scale, can be expected to give precisely the same score for each judgment (where repeat assessments are separated by a considerable number of intervening ones). A more flexible tool is thus required. We build such a tool by starting with two core assumptions:

***A*:** When a consistent assessor is presented with a set of repeat judgments, the mean of the initial set of assessments will not be significantly different from the mean score of repeat assessments.

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

Assumption B is the basis of our quality-control mechanism, and allows us to distinguish between Turkers who are working carefully and those who are merely going through the motions. We use a 100-judgment HIT structure to control same-judge repeat items and deliberately-degraded system outputs (*bad_reference* items) used for worker-intrinsic quality control (Graham et al., 2013). *bad_reference* translations for fluency are created as follows: two words in the translation are randomly selected and randomly re-inserted elsewhere in the sentence (but not as the initial or final word of the sentence).

	total wrkrs	fltrd wrkrs	Assum A holds	total segs	fltrd segs
F	557	321 (58%)	314 (98.8%)	122k	78k (64%)
A	542	283 (52%)	282 (99.6%)	102k	62k (61%)

Table 1: Total quality control filtered workers and assessments (F = fluency; A = adequacy).

Since adding duplicate words will not degrade adequacy in the same way, we use an alternate method to create *bad_reference* items for adequacy judgments: we randomly delete a short sub-string of length proportional to the length of the original translation to emulate a missing phrase. Since this is effectively a new degradation scheme, we tested against experts. For low-quality translations, deleting just two words from a long sentence often made little difference. The method we eventually settled on removed a sequence of words as a function of sentence length: $[2, 3] \rightarrow 1$, $[4, 5] \rightarrow 2$, $[6..8] \rightarrow 3$, $[9..15] \rightarrow 4$, $[16..20] \rightarrow 5$, and $[20+] \rightarrow n/5$, where n is the length of the sentence.

To filter out careless workers, scores for *bad_reference* pairs are extracted, and a difference of means test is used to calculate a worker-reliability estimate in the form of a p-value. Paired tests are then employed using the raw scores for degraded and corresponding system outputs, using a reliability significance threshold of $p < 0.05$. If a worker does not demonstrate the ability to reliably distinguish between a bad system and a better one, the judgments from that worker are discarded. This methodology means that careless workers who habitually rate translations either high or low will be detected, as well as (with high probability) those that click (perhaps via robots) randomly. It also has the advantage of not filtering out workers who are internally consistent but whose scores happen not to correspond particularly well to a set of expert assessments.

Having filtered out users who are unable to reliably distinguish between better and worse sets of translations ($p \geq 0.05$), we can now examine how well Assumption A above holds for the remaining users, the extent to which workers apply consistent scores to repeated translations. We compute mean scores for the initial and repeat items and look for even very small differences in the two distributions for each worker. Table 1 shows numbers of

workers who passed quality control, and also the percentage of reliable workers with no significant difference between mean scores for repeat items.

4 Five Years of Machine Translation

To estimate the improvement in MT that took place between 2007 and 2012, we asked workers on Amazon's Mechanical Turk (MTurk) to rate the quality of translations produced by the best-reported 2007-participating WMT system and the same for WMT 2012 (Callison-Burch et al., 2007; Callison-Burch et al., 2012). Since it is likely that over this time period, the test set has changed in difficulty, we also include in the evaluation the original test data for 2007 and 2012, translated by a single current MT system. We use the latter to calibrate the results for test set difficulty, by calculating the average difference in rating, Δ, between the 2007 and 2012 test sets. This is then added to the difference in rating for the best-reported systems in 2012 and 2007, to arrive at an overall evaluation of the 5-year gain in MT quality for a given language pair, separately for fluency and adequacy.

Experiments were carried out for German, French, Spanish into and out of English, and for Czech-to-English. English-to-Czech was omitted, because of a low response rate on MTurk. For language pairs where two systems tied for first place in the shared task, a random selection of translations from both systems was made.

To facilitate quality control, we comprise each HIT on MTurk as an assessment of 100 translations. Each individual translation is rated in isolation from other translations with workers required to iterate through 100 translations without the opportunity to revisit earlier assessments. A 100-translation HIT contains the following items:

(A) 70 randomly selected system outputs made up of roughly equal proportions of translations for each evaluated system.[3]

(B) 10 degraded versions of translations included in (A).

(C) 10 exact repeats of translations included in (A).

(D) 10 reference translations for 10 translations in (A).

[3]For instance, the current evaluation includes 4 systems, therefore (A) includes roughly 18 system outputs from each.

This results in 30 pairs of translations, of which one member is an original system output (randomly selected from (A) without replacement) paired with either a *bad_reference*, exact repeat of it, or reference translation for it, making up sets (B), (C) and (D) respectively, with a remaining 40 system outputs in (A) not paired with any other translation.

Translations are randomly ordered, but with constraints applied so that as much as possible pairs of quality control items are separated from each other. A hit of 100 translations, therefore, is divided into 10 sets of 10 translations, $S_{0..9}$, presented to the human judge in this exact order, with translations randomly jumbled only within its set, S_i. For each i in $0..4$, sets S_i and $S_{(i+5)}$ are structured as follows:

	S_i	$S_{(i+5)}$
1.	1 bad reference (from (B))	its corresponding system output (from (A))
2.	1 system output (from (C))	a repeat of it (from (A))
3.	1 reference (from (D))	its corresponding system output (from (A))
4.	1-3 above in reverse for S_i and $S_{(i+5)}$	
5.	4 system outputs (from (A))	4 system outputs (from (A))

Assessment set-up

Separate HITs were provided for evaluation of fluency and adequacy. For fluency, a single system output was displayed per screen, with a worker required to rate the fluency of a translation on a 100-point visual analog scale with no displayed point scores. A similar set-up was used for adequacy but with the addition of a reference translation (displayed in gray font to distinguish it from the system output being assessed). The Likert-type statement that framed the judgment was *Read the text below and rate it by how much you agree that*:

- [for fluency] *the text is fluent English*
- [for adequacy] *the black text adequately expresses the meaning of the gray text*.

In neither case was the source language string provided to the workers.

Tasks were published on MTurk, with no region restriction but the stipulation that only na-

tive speakers of the target language should complete HITs, and with a qualification of an MTurk prior HIT-approval rate of at least 95%. Instructions were always presented in the target language. Workers were paid US$0.50 per fluency HIT, and US$0.60 per adequacy HIT.[4]

Close to one thousand individual Turkers contributed to this experiment (some did both fluency and adequacy assessments), providing a total of more than 220,000 segments, of which 140,000 were provided by workers meeting the quality threshold.

In general, it cost approximately US$30 to assess each system, but the likelihood of it being done by a good quality worker (Table 1) effectively doubles the cost of the annotation. We rejected HITs where it was clear that random-clicking had taken place, but did not reject solely on the basis of having not met the quality control threshold, to avoid penalizing well-intentioned but low-quality workers.

Overall change in performance

Table 2 shows the overall gain made in five years. Mean scores for the two top performing systems from each shared task (BEST$_{07}$, BEST$_{12}$) are included, as well as scores for the benchmark current MT system on the two test sets (CURR$_{07}$, CURR$_{12}$). For each language pair, a HIT of 100 test segments was constructed by randomly selecting translations from the pool of $(3003+2007)\times 2$ that were available, and this results in apparently fewer assessments for the 2007 test set. In fact, numbers of evaluated translations are relative to the size of each test set. Average z scores for each system are also presented, based on the mean and standard deviation of all assessments provided by an individual worker, with positive values representing deviations above the mean of workers. In addition, we include mean BLEU (Papineni et al., 2001) and METEOR (Banerjee and Lavie, 2005) automatic scores for the same system outputs.

The CURR benchmark shows fluency scores that are 5.9 points higher on the 2007 data set than they are on the 2012 test data, with a larger difference in adequacy of 8.3 points, and hence, as expected, the 2012 test data is more challenging than 2007. Despite this, both fluency and adequacy scores for the best system in 2012 have increased by 4.5 and 2.0 points respectively, amounting to estimated average gains of 10.4 points in fluency and 10.3 points in adequacy for state-of-the-art MT on average across the seven language pairs.

Looking at the standardized scores, it is apparent that the presence of the CURR segments for the 2007 test set pushes the mean score for the 2007 best systems below zero. The presence in the HITs of reference translations also shifts standardized system evaluations below zero, because they are not attributable to any of the systems being assessed.[5]

Results for automatic metrics lead to similar conclusions: that the test set has indeed increased in difficulty; and that, in spite of this, substantial improvements have been made according to automatic metrics, +13.5 using BLEU, and +7.1 on average using METEOR.

Language pairs

Table 3 shows mean fluency and adequacy scores by language pair for translation into English. Relative gains in both adequacy and fluency for the to-English language pairs are in agreement with the estimates generated through the use of the two automatic metrics. Most notably Czech-to-English translation appears to have made substantial gains across the board achieving more than double the gain made by some of the other language pairs; results for best participating 2007 systems show that this may in part be caused by the fact that Czech-to-English translation had a lower 2007 baseline to begin with (BEST$_{07}$ F:40.8; A:41.7) in comparison to, for example, Spanish-to-English translation (BEST$_{07}$ F:56.7; A:59.0), for example.

Another notable result is that although the test data for each year's shared task is parallel across five languages, test set difficulty increases by different degrees according to human judges and automatic metrics, with BLEU scores showing substantial divergence across the to-English language pairs. Comparing BLEU scores achieved by the benchmark system for Spanish to English and Czech-to-English, for example, the benchmark system achieves close scores on the 2007 test data with a difference of only $|52.3 - 51.2| = 1.1$, compared to the score difference for the benchmark scores for translation of the 2012 test data of

[4] Since insufficient assessments were collected for French and German evaluations in the initial run, a second and ultimately third set of HITs were needed for these languages with increased payment per HIT of US$1.0 per 100-judgment adequacy HIT, US$0.65 per 100-judgment fluency HIT and later again to US$1.00 per 100-judgment fluency HIT.

[5] Scores for reference translations can optionally be omitted for score standardization.

		CURR$_{07}$	CURR$_{12}$	Δ (CURR$_{07}$ − CURR$_{12}$)	BEST$_{07}$	BEST$_{12}$	5-Year Gain (BEST$_{12}$ − BEST$_{07}$ + Δ)
fluency	score	64.1	58.2	5.9	53.5	58.0 (+4.5)	10.4
	z	0.18	0.00	0.18	−0.16	0.00 (+0.16)	0.34
	n	12,334	18,654		12,513	18,579	
adequacy	score	65.0	56.7	8.3	54.0	56.0 (+2.0)	10.3
	z	0.18	−0.07	0.25	−0.16	−0.09 (+0.07)	0.32
	n	10,022	14,870		10,049	14,979	
metrics	BLEU	41.5	30.0	11.4	25.6	27.7 (+2.1)	13.5
	METEOR	49.2	41.1	8.1	41.1	40.1 (−1.0)	7.1

Table 2: Average human evaluation results for all language pairs; mean and standardized z scores are computed in each case for n translations. In this table, and in Tables 3 and 4, all reported fluency and adequacy values are in points relative to the 100-point assessment scale.

			CURR$_{07}$	CURR$_{12}$	Δ (CURR$_{07}$ − CURR$_{12}$)	BEST$_{07}$	BEST$_{12}$	5-Year Gain (BEST$_{12}$ − BEST$_{07}$ + Δ)
DE-EN	fluency	score	65.3***	57.9	7.4	52.8	55.0* (+2.2)	9.6
		n	2,164	3,381		2,242	3,253	
	adequacy	score	63.8***	52.8	11.0	46.5	49.8** (+3.3)	14.3
		n	1,458	2,175		1,454	2,193	
	metrics	BLEU	38.3	26.5	11.8	21.1	23.8 (+2.7)	14.5
		METEOR	40.3	32.7	7.6	33.4	31.7 (−1.7)	5.9
FR-EN	fluency	score	65.9***	58.0	7.9	57.8	60.2** (+2.4)	10.3
		n	2,172	3,267		2,203	3,238	
	adequacy	score	61.0***	52.3	8.7	52.7	51.5 (−1.2)	7.5
		n	1,754	2,651		1,763	2,712	
	metrics	BLEU	39.4	32.0	7.4	28.6	31.5 (+2.9)	10.3
		METEOR	39.8	34.6	5.2	35.9	34.3 (−1.6)	3.6
ES-EN	fluency	score	68.4***	59.2	9.2	56.7	56.7 (+0.0)	9.2
		n	1,514	2,234		1,462	2,230	
	adequacy	score	68.0***	56.9	11.1	59.0***	55.7 (−3.3)	7.8
		n	1,495	2,193		1,492	2,180	
	metrics	BLEU	51.2	38.3	12.9	35.1	33.5 (−1.6)	11.3
		METEOR	45.4	37.0	8.4	39.9	36.0 (−3.9)	4.5
CS-EN	fluency	score	62.3***	49.9	12.4	40.8	50.5*** (+9.7)	22.1
		n	1,873	2,816		1,923	2,828	
	adequacy	score	62.4***	47.5	14.9	41.7	47.4*** (+5.7)	20.6
		n	1,218	1,830		1,257	1,855	
	metrics	BLEU	52.3	25.0	27.3	25.1	22.4 (−2.7)	24.6
		METEOR	44.7	31.6	13.1	34.3	30.8 (−3.5)	9.6

Table 3: Human evaluation of WMT 2007 and 2012 best systems for to-English language pairs. Mean scores are computed in each case for n translations. In this table and in Table 4, * denotes significance at $p < 0.05$; ** significance at $p < 0.01$; and *** significance at $p < 0.001$.

$|25.0 - 38.3| = 13.3$. This may indicate that the increase in test set difficulty that has taken place over the years has made the shared task disproportionately more difficult for some language pairs than for others. It does seem that some language pairs are harder to translate than others, and the differential change may be a consequence of the fact that increasing test set complexity for all languages in parallel increases the difficulty more for harder language pairs.

Table 4 shows results for translation out-of English, and once again human evaluation scores are in agreement with automatic metrics with English-to-Spanish translation achieving most substantial

			CURR$_{07}$	CURR$_{12}$	Δ (CURR$_{07}$ − CURR$_{12}$)	BEST$_{07}$	BEST$_{12}$	5-Year Gain (BEST$_{12}$ − BEST$_{07}$ + Δ)
EN-ES	fluency	score	77.2***	73.4	3.8	63.3	71.9*** (+8.6)	12.4
		n	2,286	3,318		2,336	3,420	
	adequacy	score	75.2***	68.1	7.1	62.5	67.2 (+4.7)	11.8
		n	1,410	2,039		1,399	2,112	
	metrics	BLEU	48.2	38.7	9.5	29.1	35.3 (+6.2)	15.7
		METEOR	69.9	59.6	10.3	57.0	58.1 (+1.1)	11.4
EN-FR	fluency	score	57.1	55.2	1.9	49.5	56.4 (+6.9)	8.8
		n	1,008	1,645		1,039	1,588	
	adequacy	score	64.2*	61.9	2.3	57.2	62.3 (+5.1)	7.4
		n	1,234	1,877		1,274	1,775	
	metrics	BLEU	37.2	30.8	6.4	25.3	29.9 (+4.6)	11.0
		METEOR	59.4	52.9	6.5	50.4	52.0 (+1.6)	8.1
EN-DE	fluency	score	52.3	54.1*	−1.8	53.7	55.5 (+1.8)	0.0
		n	1,317	1,993		1,308	2,022	
	adequacy	score	60.3**	57.4	2.9	58.3	58.3 (+0.0)	2.9
		n	1,453	2,105		1,410	2,152	
	metrics	BLEU	23.6	18.7	4.9	14.6	17.2 (+2.6)	7.5
		METEOR	44.7	39.1	5.6	36.7	38.0 (+1.3)	6.9

Table 4: Human evaluation of WMT 2007 and 2012 best systems for out of English language pairs. Mean scores are computed in each case for n translations.

gains for the three out-of-English language pairs, an increase of 12.4 points for fluency, and 11.8 points with respect to adequacy, while English-to-French translation achieves a gain of 8.8 for fluency and 7.4 points for adequacy. English to German translation achieves the lowest gain of all languages, with apparently no improvement in fluency, due to a negative Δ as the human fluency evaluation of the benchmark system on the expectedly easier 2007 data receives a significantly lower score compared to the benchmark system for translations of the 2012 data. This result demonstrates why fluency, the way we evaluate it without a reference translation, should not be used to evaluate MT systems without an adequacy assessment, since it is entirely possible for a low adequacy translation to achieve a high fluency score.

For all language pairs Figure 1 plots fluency, adequacy and F_1 net gain against increase in test data difficulty.

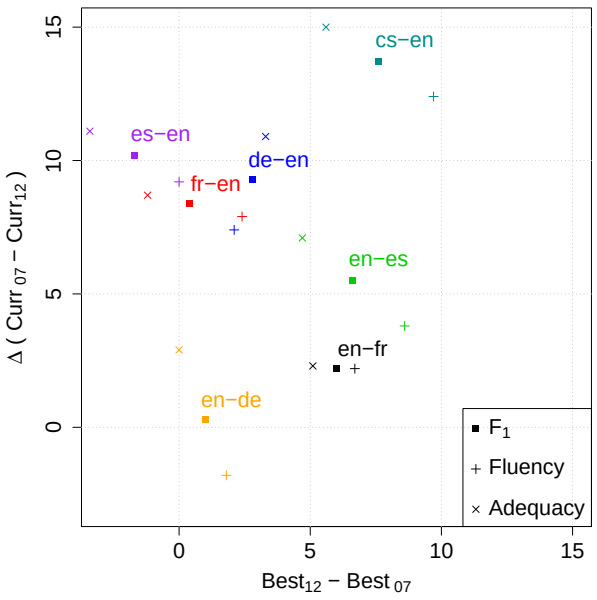

Figure 1: Mean fluency, adequacy and combined F_1 scores for language pairs.

5 Conclusion

We carried out a large-scale human evaluation of best-performing WMT 2007 and 2012 shared task systems in order to estimate the improvement made to state-of-the-art machine translation over this five year time period. Results show significant improvements have been made in machine translation of European language pairs, with Czech-to-English recording the greatest gains. It is also clear from our data that the difficulty of the task has risen over the same period, to varying degrees for individual language pairs.

We plan to make the data set publicly available.

References

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

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

S. Banerjee and A. Lavie. 2005. METEOR: An automatic metric for mt evaluation with improved correlation with human judgements. In *Proc. Wkshp. Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–73, Ann Arbor, MI.

A. Birch, B. Haddow, U. Germann, M. Nadejde, C. Buck, and P. Koehn. 2013. The feasibility of HMEANT as a human MT evaluation metric. In *Proc. 8th Wkshp. Statistical Machine Translation*, pages 52–61, Sofia, Bulgaria. ACL.

O. Bojar, M. Ercegovcevic, and M. Popel. 2011. A grain of salt for the WMT manual evaluation. In *Proc. 6th Wkshp. Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland.

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

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

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

M. Dreyer and D. Marcu. 2012. HyTER: Meaning-equivalent semantics for translation evaluation. In *Proc. 2012 Conf. North American Chapter of the ACL: Human Language Technologies*, pages 162–171, Montreal, Canada. ACL.

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

C. Lo and D. Wu. 2011. MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proc. 49th Annual Meeting of the ACL: Human Language Techologies*, pages 220–229, Portland, OR. ACL.

NIST. 2002. The 2002 nist machine translation evaluation plan. Technical report, National Institute of Standards and Technology.

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

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

M. Snover, B. Dorr, R. Scwartz, J. Makhoul, and L. Micciula. 2006. A study of translation error rate with targeted human annotation. In *Proc. 7th Biennial Conf. of the Assoc. Machine Translaiton in the Americas*, pages 223–231, Boston, MA.

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