How do Humans Evaluate Machine Translation

Francisco Guzmán **Ahmed Abdelali** **Irina Temnikova** **Hassan Sajjad** and **Stephan Vogel**
ALT Research Group
Qatar Computing Research Institute, HBKU
{fguzman,aabdelali,itemnikova,hsajjad,svogel}@qf.org.qa

Abstract

In this paper, we take a closer look at the MT evaluation process from a *glass-box* perspective using eye-tracking. We analyze two aspects of the evaluation task – the background of evaluators (*monolingual or bilingual*) and the sources of information available, and we evaluate them using time and consistency as criteria. Our findings show that *monolinguals* are slower but more consistent than *bilinguals*, especially when only target language information is available. When exposed to various sources of information, evaluators in general take more time and in the case of *monolinguals*, there is a drop in consistency. Our findings suggest that to have consistent and cost effective MT evaluations, it is better to use *monolinguals* with only target language information.

1 Introduction

Each year thousands of human judgments are used to evaluate the quality of Machine Translation (MT) systems to determine which algorithms and techniques are to be considered the new state-of-the-art. In a typical scenario human judges evaluate a system's output (or *hypothesis*) by comparing it to a source sentence and/or to a reference translation. Then, they score the hypothesis according to a set of defined criteria such as *fluency* and *adequacy* (White et al., 1994); or rank a set of hypotheses in order of preference (Vilar et al., 2007; Callison-Burch et al., 2007).

Evaluating MT output can be a challenging task for a number of reasons: it is tedious and therefore evaluators can lose interest quickly; it is complex, especially if the guidelines are not well defined; and evaluators can have difficulty distinguishing between different aspects of the translations (Callison-Burch et al., 2007). As a result, evaluations suffer from low inter- and intra-annotator agreements (Turian et al., 2003; Snover et al., 2006). Yet, as Sanders et al. (2011) argue, using human judgments is essential to the progress of MT because: (*i*) automatic translations are produced for a human audience; and (*ii*) human understanding of the *real* world allows to assess the importance of the errors made by MT systems.

Most of the research in human evaluation has focused on analyzing the criteria to use for evaluation, and has regarded the evaluation process as a *black-box*, where the inputs are different sources of information (i.e source text, reference translation, and translation hypotheses), and the output is a score (or preference ranking).

In this paper, we focus on analyzing evaluation from a different perspective. First, we regard the process as a *glass-box* and use eye-tracking to monitor the times evaluators spend digesting different sources of information (*scenarios*) before making a judgment. Secondly, we contrast how the availability of such sources can affect the outcome of the evaluation. Finally, we analyze how the background of the evaluators (in this case whether they are *monolingual* or *bilingual*) has an effect on the consistency and speed in which translations are evaluated. Our main research questions are:

- Given different *scenarios*, what source of information do evaluators use to evaluate a translation? Do they use the source text, the target text, or both? Does the availability of specific information changes the consistency of the evaluation?
- Are there differences of behavior between *bilinguals* (i.e. evaluators fluent in both source and target languages) and *monolinguals* (i.e. evaluators fluent only in the target language)? Which group is more consistent?

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 457–466,
Lisboa, Portugal, 17-18 September 2015.

Our goal is to provide actionable insights that can help to improve the process of evaluation, especially in large-scale shared-tasks such as WMT. In the next sections we summarize related work, provide details of our experimental setup, and analyze and discuss the results of our experiment.

2 Related Work

Previous work on human evaluation has focused on various aspects of the evaluation process ranging from categorization of the possible scenarios (Sanders et al., 2011) to the effectiveness of the evaluation criteria (Callison-Burch et al., 2007). Callison-Burch et al. (2007) define several criteria to evaluate the effectiveness of a MT evaluation task: (*i*) The *ease* with which humans are able to perform the task; (*ii*) the *agreement* with respect to other annotators; and (*iii*) the *speed* with which annotations can be collected.

Based on those criteria they recommended that evaluations should be done in the form of ranking translations against each other instead of assigning absolute scores to individual translation because ranking is easier to perform, can be done faster, and produces evaluations with higher levels of inter-annotator agreement. As a result, recent WMT evaluations have adopted this evaluation-by-ranking approach and instructions are *kept minimal* by only asking the evaluator to rank hypotheses from worst to best (Bojar et al., 2011).

In this work, we consider the three criteria proposed by Callison-Burch et al. (2007): *ease*, *agreement* and *speed*; but with a few differences. Regarding *ease*, instructions are kept minimal, and the evaluation criteria is left to the evaluator to decide (or discover). Furthermore, by framing the evaluation as a game we aim to keep participants engaged, and make the evaluation task easier. With respect to the other two criteria, we use them to analyze two different aspects of the evaluation process: the sources of information available to the evaluator, and the background of the evaluator.

Eye-tracking has been previously used in MT evaluation research for different purposes. Doherty et al. (2010) used eye-tracking to evaluate the comprehensibility of machine translation output in French, by asking native speakers to read MT output. They found that eye-tracking data had a slight correlation with HTER scores.

Stymne et al. (2012) applied eye-tracking to machine translation error analysis. They found that longer gaze time and and a higher number of fixations correlate with high number of errors in the MT output. Doherty and O'Brien (2014) used eye-tracking to evaluate the quality of raw machine translation output in terms of its *usability* by an end user. They concluded that eye-tracking correlates well with the other measures which they used for their study. In this work, we use eye-tracking to observe which sources of information evaluators use while performing an MT evaluation task and how this impacts the task completion time and the consistency in their judgements.

3 Method

In order to understand how humans evaluate MT, we ran an evaluation experiment using eye-tracking, involving 20 human participants, half of them *monolingual* in English and the other half *bilingual* in Spanish-English. We chose the Spanish-English language pair because of the large amount of freely available data (e.g. WMT) and the sizable pool of available participants in our environment. In our setup, we contrasted the evaluation procedure under alternative *scenarios* in which different sources of information (e.g. source sentence, reference translation) are available. To keep things simple, we only asked participants to evaluate one translation at a time and provide a single score representing the translation quality. To prevent biasing the behavior of the participants, and to encourage them to evaluate translations *naturally*, participants were not given any precise instructions regarding the requirements of a *good* translation. To increase engagement, we formulated the evaluation experiment as a game, where participants are provided feedback after each evaluation according to how close their own score was to a precomputed quality score. Below, we further describe the data used, the different *scenarios*, the background of the participants, and other details of our experiment.

3.1 Data

In our experiments we used the WMT12 (Callison-Burch et al., 2012) human evaluation data for Spanish-English systems. The data consists of 1141 ranking annotations, in which each evaluator ranked five out of the 12 participating systems.

The annotation effort generated a total of 5705 labels with an inter-annotator agreement of $\kappa = 0.222$. Unfortunately, many of the translations have rankings coming from a single evaluator only.In practical terms, this means that at least two evaluators had to evaluate the translations of the same source sentence, and at least two systems were ranked by both of those evaluators. In the WMT12 data, a total of 923 different source sentences were evaluated. From these, we kept only the 155 that complied with our requirement.

To control for length (i.e number of words), we divided the sentences into three equally sized groups based on the sentence length of their reference translations. Discarding the five longest ones the resulting sets *long*, *medium*, and *short* averaged 30.88, 18.18, and 10.18 words.

To have diversity in the quality of the translations, we collected two translations per source sentence, one of superior quality (*best*), and another one of inferior quality (*worst*). We measured quality according to the *expected wins* (Callison-Burch et al., 2012). In total, we used 300 different translations.[1]

3.2 Sources of Information

Our evaluation setup is based on a typical Appraise configuration (Federmann, 2012), where evaluators are provided with different sources of information in different areas of the screen: (*i*) the hypothesis to be evaluated; (*ii*) the source sentence; (*iii*) the context of the source sentence (previous and next sentences in the same source document); (*iv*) the reference translation for the source sentence; and (*v*) the context of the reference translation (previous and next sentences in the same reference document). Figure 1 presents a snapshot of our experimental setup, along with the labels for the corresponding areas of the screen.

To *ease* the scoring procedure, instead of providing a set of predefined levels of quality (e.g. 1 to 5), we used a continuous range (a slider from 0 to 100), and let the evaluator freely set the level of translation quality.

To contrast the effect that different sources of information have on the evaluation procedure, we explored three different evaluation *scenarios*:

- **Scenario 1** (*source-only*) shows participants the translated sentence (in English) along with the source text of the translation (in Spanish), including the context of the source sentence (one sentence before and one sentence after the translated sentence).
- **Scenario 2** (*source+target*) shows participants the translated sentence (in English), along with the source text of the translation (in Spanish), and a reference translation done by a human (also in English), plus context for both source and reference.
- **Scenario 3** (*target-only*) shows the translated sentence (in English) only with a reference translation including its context (in English).

3.3 Feedback

To keep participants engaged, they were given feedback according to a previously computed quality score for each translation. This score was calculated using a linear interpolation of the *expected wins* score obtained from the ranking evaluations (normalized to the range [0, 100]) and DISCOTK$_{party}$ (Joty et al., 2014), a high-performing automatic MT metric based on discourse (Guzmán et al., 2014), which won the WMT 2014 metrics task. This was done because expected wins only provide relative scores (i.e. which of two translations is ranked better given the same source sentence), while the participants were evaluating *absolute* scores. To keep things simple, we provided feedback based on the difference between the evaluator's score and the computed quality scores. Participants were given a five scale feedback depending on the magnitude of these differences (5: [0–10], 4: [11–20], 3: [21–30], 2: [31–40], 1: [>40]). In Section 5.2 we analyze the impact of feedback on the evaluator behavior.

3.4 Participants

In our experiment we had 20 participants 27 to 45 years old. Seven of the participants were female, and 13 were male. Seventeen of our participants were computer scientists; ten had experience with manually translating documents; and four had experience with machine translation evaluation.

All the recruited participants were proficient in English. However, half of the participants were recruited taking into account their mastery of the Spanish Language. For the analysis, participants were divided into two groups of ten people each:

[1]For reproducibility, the full data matrix can be obtained at https://github.com/Qatar-Computing-Research-Institute/wmt15eyetracking

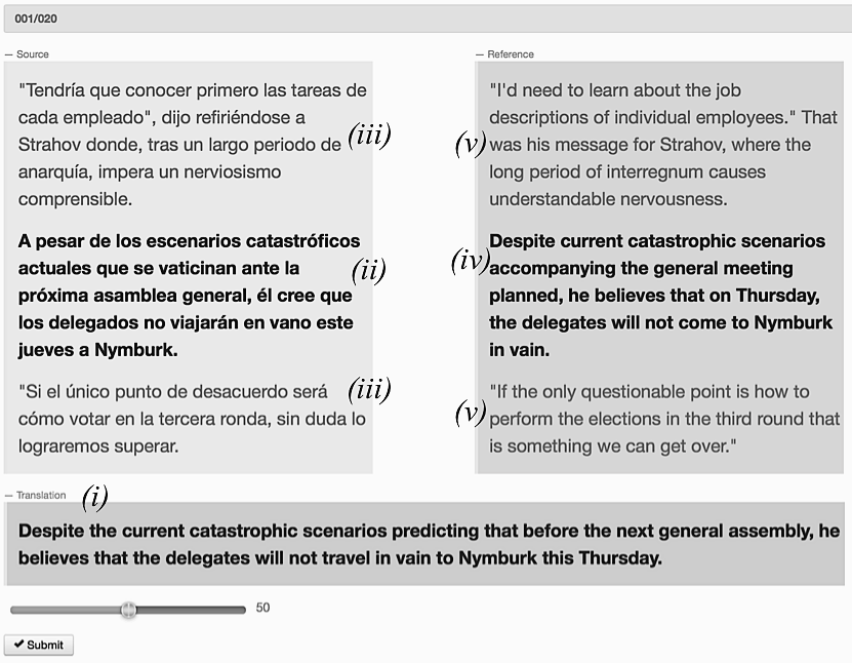

Figure 1: Our modified evaluation layout showing: the translation (*i*) ; the source (*ii*) , (*iii*) ; the reference (*iv*) , (*v*) ; and the scoring *slider*.

- **Bilingual** participants did speak the source language (Spanish) at a native or advance level of comprehension.
- **Monolingual** participants did not speak the source language. Note that this group included some speakers of other Romance languages. However, the participants insisted that their understanding of Spanish was not enough to correctly comprehend the source text.

3.5 Experimental Design

We planned our experiment to collect 1200 evaluations, 60 from each of the 20 participants. To do so, we designed an experimental matrix in which we considered the following variables: (*i*) evaluator type: *monolingual*, *bilingual*; (*ii*) length of reference: *short*, *medium*, *long*; (*iii*) *scenario*: *source-only*, *source+target*, *target-only*; and (*iv*) type of translation: *worst*, *best*.

In our experimental matrix, each participant evaluated 60 translations evenly divided into: 20 translations in each of the *scenario*; 20 translations from each length type; 30 translations of each quality type. On the other hand, each translation was evaluated by four different participants, two *bilingual* and two *monolingual*. To avoid any bias, we made sure that each evaluator saw each source sentence only once.

3.6 Eye-tracking Setup

We used the EyeTribe eye-tracker [2] to collect *gaze* information from the participants. The information was sent in messages to a modified version of Appraise[3] at a rate of 60Hz (a packet in every 16ms).

Each message contained the gaze position in the screen of both eyes, a flag indicating if the point represented a *fixation*, a time stamp, and other device-related information. To ensure optimal readings, participants were asked to calibrate the eye-tracking device before starting the experiments, and a warning message was displayed whenever the eye-tracker lost track of the participant's gaze.

3.7 Instructions and Exit Survey

Participants were asked to move as little as possible to not interrupt the readings of the eye-tracker, and to not interrupt their work while working through the translations belonging to one *scenario*, as the time for executing all sentences in one scenario was measured. Before conducting the evaluation, participants were shown two tutorials, one showing how to calibrate the eye-tracker and one showing how to conduct the experiment.

[2] http://dev.theeyetribe.com/api/

[3] Available at: https://github.com/Qatar-Computing-Research-Institute/iAppraise

After the tutorials they were asked to perform a warm-up exercise consisting of two sentences per *scenario*. Then, the participants proceeded to evaluate the 20 translations in each of the *scenarios* in the following order *source-only*, *source+target* and *target-only*[4].

After the experiment, the participants were asked to fill in an on-line exit survey, which collected their impressions about the experiment and their physiological status during the experiment.

From the survey we learned about the physiological state of the participants: 55% of them were in a normal state, 15% were slightly tired or sleepy, 25% were tired, and 10% were sleep-deprived or sick. Yet, all these reports were evenly distributed among *bilinguals* and *monolinguals*.

There were only few complaints about the setup, and they were related to: (*i*) the lack of precise instructions of what constitutes a *good* translation, (*ii*) the large range of the evaluation score (0-100), (*iii*) the difficulty to understand the context of the translations, and (*iv*) the cognitive overhead needed to evaluate *long* translations, especially in the *source+target scenario*. As expected, some of the *monolingual* participants noted that in the *source-only scenario* they mostly evaluated the readability of the translation, as they had no knowledge of the source language.

4 Results

In this section we analyze the process that participants use to evaluate translation. We focus on three different aspects. First, we use eye-tracking data to observe in which areas do participants spend most of their time. Next, we analyze the time that participants take to complete the evaluation. Finally, we analyze the scores given by the participants, and their consistency.

4.1 How Long Does it Take?

One important aspect to take into account is the time at which annotations can be collected (Callison-Burch et al., 2007). To discount the time a participant spends idle (be either by fatigue, distraction, etc.), here we analyze only the *focused* time, i.e. the amount of time a participant gaze is focused in *areas of interest*.

[4]In hindsight, randomizing the order in which the *scenarios* were performed would have allowed to answer an additional set of questions.

In our experiments, we observed that on average annotations take 26.06 seconds to be collected, which is in line with the measurements reported by Callison-Burch et al. (2007). In Table 1, we further break down the task durations by: (*i*) type of evaluator (i.e. *monolingual* and *bilingual*), (*ii*) *scenario* (i.e. *source-only*, *source+target*, and *target-only*); and (*iii*) the length of the source sentence (i.e. *short*, *medium*, *long*).

	scnr.	usr_type	*long*	*med*	*short*	avg
1	src	biling	36.89	24.54	17.92	26.46
2	src	mono	44.11	28.58	19.17	30.55
3	src+tgt	biling	40.16	23.99	15.46	26.59
4	src+tgt	mono	46.76	29.69	21.63	32.71
5	tgt	biling	26.41	15.03	10.54	**17.28**
6	tgt	mono	35.90	19.41	12.69	22.77

Table 1: Average task duration time (in seconds) according to type of setup, type of evaluator and source sentence length.

The first observation to make is that *bilingual* evaluators are consistently faster than *monolingual* evaluators in evaluation. This is true even in the *target-only* condition, where both evaluators can leverage the same amount of information (i.e. both are fluent in English). This can have two possible explanations: (*i*) *bilingual* evaluators develop *internal* rules that allow them to perform the task faster, and (*ii*) since the order of the conditions was fixed (i.e evaluators performed first the *source-only* tasks, then the *source+target* tasks and lastly the *target-only* tasks), this could mean that the *bilingual* evaluators got *more efficient* sooner, just because the *source-only* task wasn't noise to them. However, we show later that (*i*) is more plausible.

The second observation to make is that evaluators tend to take longer to evaluate scenarios with more sources of information available. This is true for *monolingual* if we analyze the results either by *scenario* or by source length[5]. Surprisingly, *monolingual* participants in the *source-only* condition perform the task 7% faster than in the *source+target* condition, which leads to hypothesize that the more information is in the screen, the longer the task will take, even if the information is not particularly useful for the task completion. On the other hand *bilingual* take the least time when evaluating *target-only* scenario.

[5]Longer source sentences have more words.

To measure the significance of our observations, we fitted a random intercepts model and analyzed the relationship between task duration time, length of the sentences, type of evaluator and type of scenario while taking into account the variability between evaluators. Therefore, as fixed effects, we had the length of the sentences, the type of evaluator (*bilingual* and *monolingual*) and the *scenario* into the model. We also included the interaction between the type of evaluator and the length of the sentences. As random effects, we had intercepts for each of the 20 evaluators. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question.

In general, the effect of *scenario* is highly significant ($\chi^2_2 = 121.71$, $p = 2.2e^{-16}$), and for long sentences the *target-only* scenario is 8.52 and 9.6 seconds faster than the *source-only* and *source+target* scenarios, respectively. The effect of the type of evaluator is also significant ($\chi^2_3 = 7.45$, $p = 0.05$), and on average *bilingual* are faster than *monolingual* by 7.76 seconds for long sentences. These results were obtained using R (R Core Team, 2015) and lme4 (Bates et al., 2015), following Winter (2013).

4.2 Where Do Evaluators Look?

The eye-tracking data allowed us to analyze the behavior of the evaluators across different conditions. In particular, we focused in the *dwell* time, i.e. the amount of time an evaluator is looking at a particular *area of interest* in the screen. In Table 2, we present the proportional *dwell* time (out of the *focused* time) that the evaluators spent in the different areas of the screen: (*i*) translation, (*ii*) source (with previous and next context), (*iii*) reference (with previous and next context), (*iv*) and the sum of the source and reference times.

From the table, the main observation is that evaluators spend most of their time looking at regions other than the translation (src+ref). This supports the hypothesis that evaluators try to understand the source and reference before making a judgment about the translation. However, there are some peculiarities worth noting. First, *bilingual* participants spend less time reading the translation than their *monolingual* counterparts.

	scnr.	usr_type	tra	ref	src	src+ref
1	src	mono	0.18	-	0.82	0.82
2	src	biling	0.12	-	0.88	0.88
3	src+tgt	mono	0.13	0.24	0.63	0.87
4	src+tgt	biling	0.07	0.16	0.78	0.93
5	tgt	mono	0.26	0.74	-	0.74
6	tgt	biling	0.19	0.81	-	0.81

Table 2: Proportional time spent by evaluators while focusing in different regions of the screen: translation (trans), reference and its context (ref), source and its context (src), and the aggregate of src and ref.

For example, this means that on average, in the *target-only* condition, a *bilingual* evaluator would spend 5 ($0.19 * 26.41$) seconds[6] focused on a *long* translation while a *monolingual* evaluator would spend 9.3 ($0.26 * 35.9$) seconds, that is almost double the time. In contrast, the difference times both *bilingual* and *monolingual* evaluators would spend reading the reference is only a factor of 1.2 (21.3 and 26.6 seconds, respectively). This tells that *bilingual* are faster (mostly) because they spend *less* time reading the translation.

Another interesting observation is that *monolingual* spend a sizable proportion of their time reading the source (which they supposedly *do not* understand), even in the *source+target* scenario. This suggests that *monolingual* evaluators develop *rules-of-thumb* to analyze the source, even if it is a foreign language (e.g. translation of named entities, numbers, dates). This can be an artifact of the relatedness between English and Spanish, or an priming effect induced by the order in which the tasks were done (i.e by asking *monolingual* evaluators to score *source-only* tasks first, we forced them into developing this behavior). The analysis of such phenomena, while interesting, is beyond the scope of this paper.

Finally, if we look across conditions, we observe that evaluators spend a larger proportion of their time evaluating the translation in the *target-only* condition than in the *source-only* and *source+target* conditions. Yet, when we calculate the expected focused time in the translation region for each condition (across different lengths and evaluator types), we obtain 4.48, 4.35 and 2.85 seconds for each condition, respectively.

[6]This time does not need to be continuously spent on the same region. For example, a evaluator might analyze a first portion of a translation, then move back to the reference, and then return to the translation.

This tells us that having more information on the screen (the case of *source+target*) decreases the total amount of time spent reading the translation. In other words, if a evaluator has more sources of information to evaluate a translation, s/he'll spend more time performing the task, but less time evaluating the translation itself.

4.3 Score Consistency

Another important aspect to take into account is how consistent are the scores provided by different evaluators, and how this consistency varies depending on the type of evaluator, and the *scenario* that is used. Unlike other studies where categorical and ordinal scores are produced, here each annotation generates a score in a continuous scale[7]. Thus, using the standard inter-annotator agreement is impractical. Instead, we evaluate *consistency* as the standard deviation of scores for each translation with respect to a class or group average (i.e. *monolingual* or *bilingual*). This quantity gives us an idea of how much variation there is in the score for a specific translation across different groups of evaluators. To be able to compare across evaluators, we normalized their individual scores to a 0-1 range using *minmax*. Then, computed the consistency as follows:

$$\sigma_c^2 = \frac{1}{N_c} \sum_{i \in T} \sum_{j \in C} (\tilde{x}_{ij} - \bar{\tilde{x}}_{ic})^2 \qquad (1)$$

where $\tilde{x}_{ij}$ is the *normalized* score of translation i by an evaluator j who belongs to class c (e.g. *monolingual*), and $\bar{\tilde{x}}_{ic}$ is the average score given to translation i by evaluators in class c, and N_c is the total number of translations scored by evaluators in class c.

In Table 3 we present the consistency measurements for *monolingual* and *bilingual* evaluators across the different conditions.

First note that *monolingual* evaluators are more consistent within their group (σ_c) than the *bilingual* evaluators. This observation holds true across all the different scenarios. Note also that *monolingual* evaluators are the *most* consistent in the *target-only* condition. We hypothesize that this is due to the longer times spent analyzing the translation in comparison to *bilingual* evaluators.

[7] Actually it is an ordinal scale from 0-100, but for practical purposes we treat it as continuous

	scnr.	usr_type	σ_c
1	src	mono	15.14
2	src	biling	16.17
3	src+tgt	mono	14.88
4	src+tgt	biling	15.96
5	tgt	mono	**14.13**
6	tgt	biling	16.81

Table 3: Consistency scores: standard deviation with respect to the class average (σ_c) for the scores produced by different types of evaluators across different conditions. Lower scores means higher consistency. Each measure is calculated over $N = 200$ points.

But also, we think this is related to the simplicity of the task. There is less information to analyze. On the other hand *bilingual*, have a larger variation, which can be attributed to the heterogeneity of *rules of thumb* that the evaluators develop from looking at the source. Finally, note how *bilingual* have a problem of consistency with the *target-only* task. Without more fine-grained information, we can only hypothesize that this is due to the lack of familiarity with the scenario. Before performing tasks in the *target-only* scenario, they were relying primarily on the source to evaluate.

4.4 Summary of Observations

We have observed that there are differences in how translations are evaluated according to the type of evaluator, and the scenario. In summary, the observations are:

- The *bilingual* evaluators perform the tasks faster than the *monolingual*. They also spend less time evaluating the translation.
- The *monolingual* evaluators are slower, but more consistent in the scores they provide.
- The more information is displayed in the screen, it will take to longer to complete the evaluation, even though, less time will be spent actually evaluating the translation. Displaying more information also correlates with lower consistency between evaluators.

5 Discussion

Using eye-tracking allowed us to dive into the process of evaluation and explore new aspects regarding the behavior of evaluators. However, there were a few additional questions that might arise from our setup and experimental results. In this section we address some of them.

5.1 Is Bilingual Adequacy Necessary?

Bilingual evaluators are considered to be the *gold standard* for the evaluation of machine translation (Dorr et al., 2011). However, the use of monolingual evaluators has been previously advocated, since the end-users of MT are in fact monolingual (Sanders et al., 2011). The results obtained in this paper lead us to challenge the inclusion of *bilingual* evaluators for MT evaluation. As seen in the results, *monolingual* evaluators were slower than *bilinguals*, but they were more consistent in their evaluations. Given the open-ended nature of *bilingual* evaluation (e.g. given a source text, they can formulate their own set of plausible translations), we believe that the evaluations of *bilinguals* can be more subjective and prone to influence by the evaluator's background and knowledge of a specific subject. Moreover, recruiting *bilingual* evaluators can be harder and more expensive. We consider that consistency should be a primary goal of any evaluation task. Therefore, it seems more practical to rely only on *monolinguals* for the evaluation of machine translation. Our findings are in line with the observations in the post-editing community where *monolinguals* were more apt for the task and improved the fluency and comprehensibility of translations (Mitchell et al., 2013). Our findings are also in partial agreement with White et al. (1993) (which is not directly comparable to our work, as it does not compare monolinguals and bilinguals performing the same task), who state that less time is spent in evaluation techniques that use only target side information.

5.2 Can Feedback Bias the Evaluation?

The process of evaluation can be cumbersome, especially if the evaluation sessions last for long; hence we used feedback to boost the engagement of participants throughout the evaluation process. This is a double-edged sword, as the feedback has the potential to bias the evaluators and influence their decision.

To rule-out any potential bias from the feedback, we investigated the effects that the progression in which the tasks were performed might have on the differences between the evaluator scores and the feedback scores.

If the evaluators *learned* to reproduce the feedback scores, we would expect that the feedback error (τ_c) would decrease as a function of time. We calculated the feedback error as follows:

$$\tau_c^2 = \frac{1}{N_c} \sum_{i \in T} \sum_{j \in C} (\tilde{x}_{ij} - f_i)^2 \qquad (2)$$

where f_i is the feedback score for translation i, and other variables are the same as in eq. 1.

We fitted a linear model to the data, using the *scenario*, the evaluator type and the progression (time) as predictors; and the feedback error as a response. We did not find that the progression had any significant effect ($p = 0.2856$) on the feedback error. This means that the feedback did not bias the scoring behavior of the evaluators.

5.3 Can We do More with Eye-tracking?

Eye-tracking technology has proven useful in different scenarios related to translation. Yet, here we have only used the eye-tracking device to measure the *dwell* time an evaluator spends reading a specific portion of the screen. Nonetheless, one can think of more refined uses for this technology.

Potentially, using eye-tracking can give us a fine-grained insight on how evaluators differentiate *good* from *bad* translations, making it easier to *learn* the intrinsic rules of thumb that they use during the evaluation process. The applications for this are manifold. For example, by learning which type of errors (e.g. morphological, syntactic, semantic) can make a stronger impact on the reading behavior while evaluating, we could help to develop *better* automatic MT evaluation metrics. Additionally, we can use gaze-data to model the evaluation score (or rank) given by an evaluator, and thus reduce the subjective score bias. This can help to alleviate the high variance found in evaluation.

However, there are several challenges that need to be solved before moving forward in this nascent area. The most important is related to the accuracy of the eye-tracking devices, which is a requirement to track which specific words are looked-at in the screen.

Eye-tracking errors can be divided into two categories: variable (device-related precision) and systematic. Fortunately, the former has improved over the past years, and high-precision devices can be now acquired for only a few hundred dollars. The latter, however is more complex. Often, a loss in accuracy known as *drift* is observed as time progresses, requiring frequent re-calibrations of the eye-tracking device.
This can be due to evaluator movements, and other environmental factors. Reducing and eliminating drift is imperative to make progress in this area. Up to now, only heuristic approaches have been proposed (Mishra et al., 2012), leaving plenty of room for improvement.

6 Conclusion

In this paper, we analyzed the process of MT evaluation from a *glass-box* perspective, using eye-tracking data. We contrasted two main aspects of the evaluation tasks: the background of the evaluators, and the sources of information available to them during the evaluation task. We used time and consistency as our main criteria for comparison. Our results show that: (*i*) *monolingual* evaluators take relatively longer to evaluate translations (except when only the target language information is available, then they complete the tasks in less time), yet they are more consistent in their judgements. (*ii*) The amount of information provided to evaluators can affect their performance. We observed that when more information is available, the tasks take longer to complete, and yield less consistent results.

Therefore, based on our empirical results, we suggest that future evaluation campaigns be done with *monolingual* evaluators in a *target-only scenario*. We argue that this setting can increase the consistency of results while reducing the potential costs of recruiting *bilinguals*.

In future studies we would like to extend our explorations into using eye-tracking to model the behavior of evaluators and to help predict reliable and unreliable translations. In particular, we would like to explore the application of eye-tracking in ranking scenarios. We believe that given the popularity and availability of *low-cost* devices, eye-tracking can position itself as a useful aid to reduce subjectivity in evaluation.

References

Douglas Bates, Martin Maechler, Benjamin M. Bolker, and Steven Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, arXiv:1406.5823.

Ondrej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A grain of salt for the wmt manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, Scotland.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada, June. Association for Computational Linguistics.

Stephen Doherty and Sharon O'Brien. 2014. Assessing the Usability of Raw Machine Translated Output: A User-Centered Study Using Eye Tracking. *International Journal of Human-Computer Interaction*, 30(1):40–51.

Stephen Doherty, Sharon O'Brien, and Michael Carl. 2010. Eye Tracking as an MT Evaluation Technique. *Machine translation*, 24(1):1–13.

Bonnie Dorr, Matthew Snover, and Nitin Madnani. 2011. Introduction. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation*, pages 745–758. Springer.

Christian Federmann. 2012. Appraise: An Open-source Toolkit for Manual Evaluation of MT Output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, USA.

Shafiq Joty, Francisco Guzmán, Lluís Màrquez, and Preslav Nakov. 2014. DiscoTK: Using discourse structure for machine translation evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA.

Abhijit Mishra, Michael Carl, and Pushpak Bhattacharyya. 2012. A Heuristic-Based Approach for Systematic Error Correction of Gaze Data for Reading. In *Proceedings of the First Workshop on Eye-tracking and Natural Language Processing*, Mumbai, India.

Linda Mitchell, Johann Roturier, and Sharon O'Brien. 2013. Community-based Post-editing of Machine-translated Content: Monolingual vs. Bilingual. In *Proceedings of the Machine Translation Summit XIV*, Nice, France.

R Core Team, 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Gregory Sanders, Mark Przybocki, Nitin Madnani, and Matthew Snover. 2011. Human Subjective Judgments. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation*, pages 750–759. Springer.

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 the 7th Biennial Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA.

Sara Stymne, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull, and Martin Wester. 2012. Eye Tracking as a Tool for Machine Translation Error Analysis. In *Proceedings of the International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Joseph Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. In *Proceedings of Machine Translation Summit IX*, New Orleans, LA, USA.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E. Banchs. 2007. Human Evaluation of Machine Translation Through Binary System Comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

John S White, Theresa A O'Connell, and Lynn M Carlson. 1993. Evaluation of machine translation. In *Proceedings of the workshop on Human Language Technology*, Stroudsburg, PA, USA.

John White, Theresa O'Connell, and Francis O'Mara. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings of the Association for Machine Translation in the Americas Conference*, Columbia, Maryland, USA.

Bodo Winter. 2013. Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499.