Reading metrics for estimating task efficiency with MT output

Sigrid Klerke[†] **Sheila Castilho**[*] **Maria Barrett**[†] **Anders Søgaard**[†]
[†]CST, University of Copenhagen, Denmark
{skl, barrett, soegaard}@hum.ku.dk
[*]CNGL/SALIS, Dublin City University, Ireland
castils3@mail.dcu.ie

Abstract

We show that metrics derived from recording gaze while reading, are better proxies for machine translation quality than automated metrics. With reliable eye-tracking technologies becoming available for home computers and mobile devices, such metrics are readily available even in the absence of representative held-out human translations. In other words, reading-derived MT metrics offer a way of getting cheap, online feedback for MT system adaptation.

1 Introduction

What's a good translation? One way of thinking about this question is in terms of what the translations can be used for. In the words of Doyon et al. (1999), "a poor translation may suffice to determine the general topic of a text, but may not permit accurate identification of participants or the specific event." Text-based tasks can thus be ordered according to their tolerance of translation errors, as determined by actual task outcomes, and task outcome can in turn be used to measure the quality of translation (Doyon et al., 1999).

Machine translation (MT) evaluation metrics must be both adequate and practical. Human task performance, say participants' ability to extract information from translations, is perhaps the most adequate measure of translation quality. Participants' direct judgements of translation quality may be heavily biased by perceived grammaticality and subjective factors, whereas task performance directly measures the usefulness of a translation. Of course different tasks rely on different aspects of texts, but some texts are written with a single purpose in mind.

In this paper, we focus on logic puzzles. The obvious task in logic puzzles is whether readers can solve the puzzles when given a more or less erroneous translation of the puzzle. We assume task performance on logic puzzles is an adequate measure of translation quality *for logic puzzles*.

Task-performance is not always a practical measure, however. Human judgments, whether from direct judgments or from answering text-related questions, takes time and requires recruiting and paying individuals. In this paper, we propose various metrics derived from natural reading behavior as proxies of task-performance. Reading has several advantages over other human judgments: It is fast, is relatively unbiased, and, most importantly, something that most of us do effortlessly all the time. Hence, with the development of robust eye tracking methods for home computers and mobile devices, this can potentially provide us with large-scale, on-line evaluation of MT output.

This paper shows that reading-derived metrics are better proxies of task-performance than the standard automatic metric BLEU. Note also that on-line evaluation with BLEU is biased by what held-out human translations you have available, whereas reading-derived metrics can be used for tuning systems to new domains and new text types.

In our experiments, we include simplifications of logic puzzles and machine translations thereof. Our experiments show, as a side result, that a promising approach to optimizing machine translation for task performance is using text simplification for pre-processing the source texts. The intuition is that translation noise is more likely to make processing harder in more complex texts.

Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning, pages 6–13,
Lisbon, Portugal, 18 September 2015.

1.1 Contributions

- We present an experimental eye-tracking study of 20 participants reading simplifications and human/machine translations of 80 logic puzzles.[1]
- This is, to the best of our knowledge, the first study to correlate reading-derived metrics, human judgments and BLEU with task performance for evaluating MT. We show that human judgments do not correlate with task performance. We also show that reading-derived metrics correlate significantly with task performance ($-.36 < r < -.35$), while BLEU does not.
- Finally, our results suggest that practical MT can benefit much from incorporating sentence compression or text simplification as a pre-processing step.

2 Summary of the experiment

In our experiments, we presented participants with 80 different logic puzzles and asked them to solve and judge the puzzles while their eye movements were recorded. Each puzzle was edited into five different versions: the original version in English (L2), a human simplification thereof (S(·)), a human translation into Danish (L1) and a machine translation of the original (M(·)), as well as a machine translation of the simplification (M(S(·))). Consequently, we used 400 different stimuli in our experiments. The participants were 20 native speakers of Danish with proficiency in English.

We record fixation count, reading speed and regression proportion (amount of fixations landing on previously read text) from the gaze data. Increased attention in the form of reading time and re-reading of previously read text are well-established indicators of increased cognitive processing load, and they correlate with typical readability indicators like word frequency, length and some complex syntactic structures (Rayner et al., 2013; Rayner, 1998; Holmqvist et al., 2011). We study how these measures correlate with MT quality, as reflected by human judgments and participants' task performance.

We thereby assume that the chance of quickly solving a task decreases when more resources are required for understanding the task. By keeping the task constant, we can assess the relative impact of the linguistic quality of the task formulation. We hypothesise that our five text versions (L1, L2, M(·), S(·), M(S(·))), can be ranked in terms of processing ease, with greater processing ease allowing for more efficient task solving.

The experiments are designed to test the following hypothesized partial ordering of the text versions (summarized in Table 1): text simplification (S(·)) eases reading processing relative to second language reading processing (L2) while professional human translations into L1 eases processing more (**H1**). In addition, machine translated text (M(·)) is expected to ease the processing load, but less so than machine translation of sim-

Math
A DVD player with a list price of $100 is marked down 30%. If John gets an employee discount of 20% off the sale price, how much does John pay for the DVD player?
1: 86.00
2: 77.60
3: 56.00
4: 50.00

Conclude
Erin is twelve years old. For three years, she has been asking her parents for a dog. Her parents have told her that they believe a dog would not be happy in an apartment, but they have given her permission to have a bird. Erin has not yet decided what kind of bird she would like to have.
Choose the statement that logically follows
1: Erin's parents like birds better than they like dogs.
2: Erin does not like birds.
3: Erin and her parents live in an apartment.
4: Erin and her parents would like to move.

Evaluate
Blueberries cost more than strawberries.
Blueberries cost less than raspberries.
Raspberries cost more than both strawberries and blueberries.
If the first two statements are true, the third statement is:
1: TRUE
2: FALSE
3: Impossible to determine

Infer
Of all the chores Michael had around the house, it was his least favorite. Folding the laundry was fine, doing the dishes, that was all right. But he could not stand hauling the large bags over to the giant silver canisters. He hated the smell and the possibility of rats. It was disgusting.
This paragraph best supports the statement that:
1: Michael hates folding the laundry.
2: Michael hates doing the dishes.
3: Michael hates taking out the garbage.
4: Michael hates cleaning his room.

Figure 1: Logic puzzles of four categories. The stimuli contain 20 of each puzzle category.

[1]The data will be made available from https://github.com/coastalcph

H1:	L1	$\prec$ S($\cdot$) $\prec$	L2
H2:	L1	$\prec$ M(S($\cdot$)) $\prec$ M($\cdot$) $\prec$	L2

Table 1: Expected relative difficulty of processing. L1 and L2 are human edited texts in the participants' native and non-native language, respectively, S($\cdot$) are manually simplified texts, M($\cdot$) are machine translated texts and M(S($\cdot$)) are machine translations of manually simplified texts.

plified text (M(S($\cdot$))), although both of these machine translated versions are still expected to be more demanding than the professionally translated original text (L1). Table 1 provides an overview of the hypotheses and the expected relative difficulty of processing each text version.

2.1 Summary of the findings

Our experimental findings are summarized as follows: The data supports the base assumption that L1 is easier than L2. We only find *partial* support for H1; While S($\cdot$) tends to be easier to comprehend than L2, also leading to improved task performance, S($\cdot$) is ranked as easier to process than L1 as often as the opposite, hypothesised ranking. This indicates that our proficient L2 readers may be benefitting as much from simplification as from translation in reasoning tasks. We also only find *partial* support for H2: The relative ordering of the human translations, L1, and the two machine translated versions, M(S($\cdot$)) and M($\cdot$), is supported and we find that the simplification improves MT a lot with respect to reading processing. However, participants tended to perform better with the original L2 logic puzzles compared to the machine translated versions. In other words, MT hurts while both manual simplification and translation help even proficient L2 readers. In sum, simplification seems necessary if L2-to-L1 MT is to ease comprehension, and not make understanding harder for readers with a certain L2 command level.

Importantly, we proceed to study the correlation of our eye-tracking measures, human judgments and BLEU (Papineni et al., 2002) with task performance. There has been considerable work on how various automatic metrics correlate with human judgments, as well as on inter-annotator consistency among humans judging the quality of translations (Callison-Burch et al., 2008). Various metrics have been proposed over the years, but BLEU (Papineni et al., 2002) remains the *de facto* state-of-the-art evaluation metric. Our findings, related to evaluation, are, as already mentioned, that (a) human judgments surprisingly do not correlate with task performance, and that (b) the reading-derived metrics TIME and FIXATIONS correlate strongly with task performance, while BLEU does not. This, in our view, questions the validity of human judgments and the BLEU metric and shows that reading-derived MT metrics may provide a better feedback in system development and adaptation.

3 Detailed description of the experiment

3.1 Stimuli

In this section, we describe the texts we have used for stimuli, as well as the experimental design and our participants.

We selected a set of 80 logic puzzles written in English, all with multiple-choice answers.[2] The most important selection criterium was that participants have to reason about the text and cannot simply recognize a few entities directly to guess the answer. The puzzles were of four different categories, all designed to train logic reasoning and math skills in an educational context. We chose 20 of each of the four puzzle categories to ensure a wide variety of reasoning requirements. Figure 1 shows an example question from each category.

The English (L2) questions and multiple choice answer options were translated into Danish (L1) by professional translators. The *question text* was manually simplified by the lead author (S($\cdot$)). Both of the English versions were machine-translated into Danish (M($\cdot$), M(S($\cdot$))).[3] This results in the five versions of the question texts, which were used for analysis. The multiple-choice answer options were not simplified or machine translated. Thus the participants saw either the original English answers or the human-translated Danish answers, matching the language of the question text. The average number of words and long words in each of the five versions are reported in Table 2.

Simplification is not a well-defined task and is often biased intentionally to fit a target audience or task. To allow for comparison with parallel simplification corpora, we classified the applied simplification operations into the following set of seven abstract simplification operations

[2] From LearningExpress (2005).

[3] Google Translate, accessed on 29/09/2014 23.33 CET.

Variant	# Long words mean	std	# Words mean	std
L2	9.56	6.67	38.33	19.29
S(·)	8.78	5.90	35.78	17.43
L1	10.22	6.97	38.87	21.28
M(S(·))	9.70	6.75	35.19	19.07
M(·)	10.35	6.74	36.53	19.04

Table 2: Mean and standard deviation of number of words and number of words with more than seven letters per question for all five versions.

Simplification	%
Lexical substitution	27.4
Paraphrase	24.2
Deletion	23.1
Information reordering	11.3
Anaphora substitution	7.5
Discourse marker insertion	4.3
Sentence splitting	2.2

Table 3: Simplification operations (SOps). The total number of applied SOps was 186, the average number of SOps applied per question was 2.0 (std 1.3).

and present their relative proportion in Table 3: Sentence splitting. information deletion and information reordering, discourse marker insertion (e.g., *and*, *but*), anaphora substitution (e.g., *Zoe's garden* vs. *the garden*), other lexical substitutions (e.g., *dogwoods* vs. *dogwood trees*) and paraphrasing (e.g., *all dogwoods* vs. *all kinds of dogwood trees*). On average 2.0 simplification operations was performed per question, while a total of 28.7% of the questions were left unchanged during simplification. All simplified questions still required the reader to understand and reason about the text. The simplifications were performed with the multiple answer texts in mind; leaving any information referenced in the answers intact in the question, even when deleting it would have simplified the question text.

3.2 Experimental design

The experiment followed a Latin-square design where each participant completed 40 trials, judging and solving 40 different puzzles, eight of each of the five versions.

A trial consisted of three tasks (see Figure 2): a comprehension task, a solving task and a comparison task. Each trial was preceded by a 1.5 second display of a fixation cross. The remainder of the trial was self-paced. During the entire trial - i.e., for the duration of the three tasks - the question text was presented on the top part of the screen. In the comprehension task, the participant was asked to rate the comprehensibility of the question text on a 7-point Likert scale that was presented at the bottom part of the screen. This score is called COMPREHENSION, henceforth. This is our rough equivalent of human judgments of translation quality. For the solving task, the multiple-choice answer options was presented in the middle part of the screen below the question text and the participant indicated an answer or "don't know" option in the bottom part of the screen. The measure EFFICIENCY, which was also introduced in Doherty and O'Brien (2014), is the number of correct answers given for a version, C_v over the time spent reading and solving the puzzles of that version, S_v: $E = \frac{C_v}{S_v}$. This score is our benchmarking metric below.

In the last task, COMPARISON, a different version of the same question text was presented below the first question text, always in the same language. Participants were asked to assess which version provided a better basis for solving the task using a 7-point Likert scale with a neutral midpoint. The three leftmost options favored the text at the top of the screen, while the three rightmost choices favored the text at the lower half of the screen.

Each participant completed three demo trials with the experimenter present. Participants were kept naïve with regards to the machine translation aspect of the study. They were instructed to solve the puzzles as quickly and accurately as possible and to judge COMPREHENSION and COMPARISON quickly. Each session included a 5-10 minute break with refreshments halfway through. At the end of the experiment a brief questionnaire was completed verbally. All participants completed the entire session in 70–90 minutes.[4]

3.2.1 Apparatus

The stimuli were presented in black letters in the typeface Verdana with a letter size of 20 pixels (ca. .4° visual angle) on a light gray background with 100 pixels margins. The eye tracker was a Tobii

[4]Participants received a voucher for 10 cups of tea/coffee upon completion.

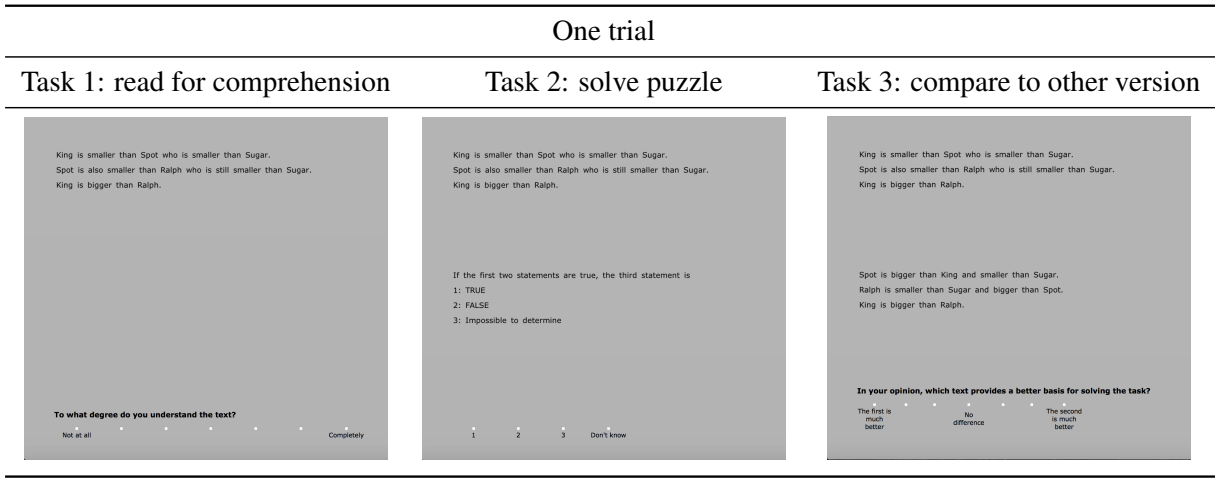

Figure 2: Illustration of one trial. Each trial consists of three individual tasks. The top third of the screen displays the target text and is fixed for the duration of the entire trial.

X120, recording both eyes with 120hz sampling rate. We used Tobii Studio standard settings for fixation detection. The stimuli was presented on a 19" display with a resolution of 1920 x 1080 pixels and a viewing distance of ca 65 cm. Here we focus on the initial reading task and report total reading time per word (TIME), number of fixations per word (FIXATIONS) and proportion of regressions (REGRESSIONS). The calculations of the eyetracking measures are detailed in Section 4.3.

3.2.2 Participants

We recruited participants until we obtained a total of 20 recordings of acceptable quality. In this process we discarded two participants due to sampling loss. Another two participants were dismissed due to unsuccessful calibration. All participants completed a pre-test questionnaire identifying themselves as native Danish speakers with at least a limited working proficiency of English. None of the participants had been diagnosed with dyslexia, and all had normal or corrected to normal vision. The 20 participants (4 males) were between 20 and 34 years old (mean 25.8) and minimum education level was ongoing bachelor's studies.

4 Results

The mean values for all metrics and the derived rankings of the five versions are presented in Table 4. Significance is computed using Student's paired t-test, comparing each version to the version with the largest measured value. Table 5 presents correlations with task performance (EFFICIENCY) for each measure. We describe the correlations, and their proposed interpretation, in Section 4.4.

4.1 Subjective measures

We elicited subjective evaluations of text comprehension and pairwise comparisons of versions' usefulness for solving the puzzles. Note that participants evaluate MT output significantly lower than human-edited versions.

We treated the pairwise COMPARISON scores as votes, counting the preference of one version as equally many positive and negative votes on the preferred version and the dis-preferred version, respectively. With this setup, we maintain zero as a neutral evaluation. COMPARISON was only made within the same language, so the scores should not be interpreted across languages. Note, however, how COMPARISON results show a clear ranking of versions within each language.

4.2 Task performance measures

The task performance is reported as the EFFICIENCY, i.e., correct answers per minute spent reading and solving puzzles. We observe that the absolute performance ranges from 48% to 52% correct answers. This is well above chance level (27%), and does not differ significantly between the five versions, reflecting that the between-puzzles difference in difficulty level, as expected, is much larger than the between-versions difference.

EFFICIENCY, however, reveals a clearer ranking. Participants were less efficient solving logic

VERSION	L1	M(S(·))	μ M(·)	S(·)	L2	RANKINGS
COMPREHENSION	5.58	**4.51	**4.50	5.61	5.46	S(·) ≺ L1 ≺ L2 ≺ M(S(·)) ≺ M(·)
COMPARISON	1.62	**−.54	**−1.07	.43	**−.43	L1 ≺ M(S(·)) ≺ M(·) | S(·) ≺ L2
EFFICIENCY	.94	.90	**0.80	1.0	.87	S(·) ≺ L1 ≺ M(S(·)) ≺ L2 ≺ M(·)
TIME	.54	.62	.65	.55	.54	L1 ≺ L2 ≺ S(·) ≺ M(S(·)) ≺ M(·)
REGRESSIONS	15.59	16.49	16.78	13.76	14.40	S(·) ≺ L2 ≺ L1 ≺ M(S(·)) ≺ M(·)
REGRESSIONS	17.77	18.46	19.15	15.55	16.55	S(·) ≺ L2 ≺ L1 ≺ M(S(·)) ≺ M(·)

Table 4: Mean values for the five text versions. COMPREHENSION and COMPARISON are Likert scale scores respectively ranging from 0 to 7 and from -3 to 3, EFFICIENCY is correct answers relative to reading speed, TIME is seconds per word, FIXATIONS is number of fixations per word and REGRESSIONS is proportion of re-fixations (**: Student's paired t-test relative to largest mean value $p < 0.001$)

puzzles when presented with machine translations of the original puzzles. The machine translations of the simplified puzzles actually seemingly *eased* task performance, compared to using the English originals, but differences are not statistically significant. The simplified English puzzles led to the best task performance.

4.3 Eye-tracking measures

The reading times in seconds per word (TIME) are averages over reading times while fixating at the question text located on the upper part of the screen during the first sub-task of each trial (judging comprehension). This measure is comparable to normalized total reading time in related work. Participants spent most time on the machine translations, whether of the original texts or the simplified versions.

The measure FIXATIONS similarly was recorded on the question part of the text during the initial comprehension task, normalized by text length, and averaged over participants and versions. Again we observe a tendency towards more fixations on machine translated text, and fewest on the human translations into Danish.

Finally, we calculated REGRESSIONS during initial reading as the proportion of fixations from *the furthest word read* to a preceding point in the text. Regressions may indicate confusion and on average account for 10-15% of fixations during reading (Rayner, 1998). Again we see more regressions with machine translated text, and fewest with simplified English puzzles.

4.4 Correlations between measures

We observe the following correlations between our measures. All correlations with EFFICIENCY are shown in Table 5. First of all, we found no correlations between subjective measures and eye-tracking measures nor between subjective measures and task performance. The two subjective measures, however, show a strong correlation (Spearman's $r = .50$ $p < .001$). EFFICIENCY shows significant negative correlation with both of the eye-tracking measures TIME (Pearson's $r = -.35$ $p < .001$ and FIXATIONS (Pearson's $r = -.36$ $p < .001$), but not REGRESSIONS . Within the group of eye-tracking measures TIME and FIXATION exhibit a high correlation ($r = 0.94$ $p < .001$). REGRESSIONS is significantly negatively correlated with both of these (Pearson's $r = -.38$ $p < .001$ and Pearson's $r = -.43$ $p < .001$, respectively).

We obtain BLEU scores (Papineni et al., 2002)

	Data used	r	$p \leq .001$
COMPREHENSION	all	.25	-
	M(S(·))	.36	-
	M(·)	-.27	-
COMPARISON	all	.13	-
	M(S(·))	.06	-
	M(·)	.26	-
TIME	all	-.35	✓
	M(S(·))	-.19	-
	M(·)	-.54	-
FIXATIONS	all	-.36	✓
	M(S(·))	-.26	-
	M(·)	-.57	-
REGRESSIONS	all	-.17	-
	M(S(·))	.01	-
	M(·)	-.33	-
BLEU	M(S(·))	-.13	-
	M(·)	-.17	-

Table 5: Correlations with EFFICIENCY (Pearson's r). BLEU only available on translated text. Correlation reported on these subsets for comparability.

by using the human-translated Danish text (L1) as reference for both of the MT outputs, M(·) and M(S(·)). The overall BLEU score for M(·) version is .691, which is generally considered very good, and .670 for M(S(·)). The difference is not surprising, since M(S(·)) inputs a different (simpler) text to the MT system. On the other hand, given that our participants tended to be more efficiently comprehending and solving the logic puzzles using M(S(·)), this already indicates that BLEU is not a good metric for talking about the usefulness of translations of instructional texts such as logic puzzles.

Our most important finding is that BLEU does not correlate with EFFICIENCY, while two of our reading-derived metrics do. In other words, the normalised reading time and fixation counts are better measures of task performance, and thereby of translation quality, than the state-of-the-art metric, BLEU in this context. This is an important finding since reading-derived metrics are potentially also more useful as they do not depend on the availability of professional translators.

5 Discussion

Several of our hypotheses were in part falsified. L2 is solved more efficiently by our participants than M(·), not the other way around. Also, M(S(·)) is judged as harder to comprehend than S(·) and consistently ranked so by all metrics. These observations suggest that MT is not assisting our participants despite the fact that L2 ranks lower than L1 in four out of five comparisons. Our participants are university students and did not report to have skipped any questions due to the English text suggesting generally very good L2 skills.

If we assume that EFFICIENCY – as a measure of task performance – is a good measure of translation quality (or usefulness), we see that the best indicator of translation quality that only takes the initial reading into account are FIXATIONS and TIME. This indicates that FIXATIONS and TIME may be better MT benchmarking metrics than BLEU.

6 Related work

Eye tracking has been used for MT evaluation in both post-editing and instruction tasks (Castilho et al., 2014; Doherty and O'Brien, 2014).

Doherty et al. (2010) also used eye-tracking measures for evaluating MT output and found fixation count and gaze time to correlate negatively with binary quality judgments for translation segments, whereas average fixation duration and pupil dilation were not found to vary reliably with the experimental conditions. A notable shortcoming of that study is that the translated segments in each category were different, making it impossible to rule out that the observed variation in both text quality and cognitive load was caused in part by an underlying variation in content complexity.

This shortcoming was alleviated in a recent reanalysis of previous experiments (Doherty and O'Brien, 2014; Doherty et al., 2012) which compares the usability of raw machine translation output in different languages and the original, well-formed English input. In order to test usability, a plausible task has to be set up. In this study the authors used an instructional text on how to complete a sequence of steps using a software service, previously unknown to the participants. MT output was obtained for four different languages and three to four native speakers worked with each output. Participants' subjective assessment of the usability of the instructions, their performance in terms of efficiency and the cognitive load they encountered as measured from eye movements were compared across languages. The results of this study supports the previous finding that fixation count and total task time depends on whether the reader worked with the original or MT output, at least when the quality of the MT output is low. In addition, goal completion and efficiency (total task time relative to goal completion) as well as the number of shifts (between instructions and task performance area) were shown to co-vary with the text quality.

Castilho et al. (2014) employed a similar design to compare the usability of lightly post-edited MT output to raw MT output and found that also light post-editing was accompanied by fewer fixations and lower total fixation time (proportional to total task time) as well as fewer attentional shifts and increased efficiency.

In contrast, Stymne et al. (2012) found no significant differences in total fixation counts and overall gaze time (proportional to total task time), when directly comparing output of different MT systems with expected quality differences. However, they showed that both of these two eye-tracking measures were increased for the parts of the text containing errors in comparison with

error-free passages. In addition, they found gaze time to vary with specific error types in machine translated text.

From an application perspective, Specia (2011) suggested the time-to-edit measure as an objective and accessible measure of translation quality. In their study it outperformed subjective quality assessments as annotations for a model for translation candidate ranking. Their tool was aimed at optimizing the productivity in post-editing tasks.

Eye tracking can be seen as a similarly objective metric for fluency estimation (Stymne et al., 2012). The fact that eye tracking does not rely on translators makes annotation even more accessible.

Both Doherty and O'Brien (2014) and Castilho et al. (2014) found subjective comprehensibility, satisfaction and likelihood to recommend a product to be especially sensitive to whether the instructional text for the product was raw MT output. This suggests that the lower reliability of subjective evaluations as annotations could be due to a bias against MT-specific errors. Only Stymne et al. (2012) report the correlations between eye movement measures and subjective assessments and found only moderate correlations.

This work is to the best of our knowledge the first to study the correlation of reading-derived MT metrics and task performance. Since we believe task performance to be a more adequate measure of translation quality – especially when the texts are designed with a specific task in mind – we therefore believe this to be a more adequate study of the usefulness of reading-derived MT metrics than previous work.

7 Conclusion

We presented an eye-tracking study of participants reading original, simplified, and human/machine translated logic puzzles. Our analysis shows that the reading-derived metrics TIME and FIXATIONS obtained from eye-tracking recordings can be used to assess translation quality. In fact, such metrics seem to be much better proxies of task performance, i.e., the practical usefulness of translations, than the state-of-the-art quality metric, BLEU.

References

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*, pages 70–106. Association for Computational Linguistics.

Sheila Castilho, Sharon O'Brien, Fabio Alves, and Morgan O'Brien. 2014. Does post-editing increase usability? a study with Brazilian Portuguese as target language. In *EAMT*.

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.

Stephen Doherty, Dorothy Kenny, and Andrew Way. 2012. A user-based usability assessment of raw machine translated technical instructions. In *AMTA*.

Jennifer Doyon, Kathryn B Taylor, and John S White. 1999. Task-based evaluation for machine translation. In *Proceedings of Machine Translation Summit VII*, volume 99.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

LearningExpress. 2005. *501 Challenging Logic and Reasoning Problems*. 501 Series. LearningExpress.

Kishore Papineni, Salim Roukus, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Keith Rayner, Alexander Pollatsek, and D Reisberg. 2013. Basic processes in reading. *The Oxford Handbook of Cognitive Psychology*, pages 442–461.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Lucia Specia. 2011. Exploiting objective annotations for measuring translation post-editing effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 73–80.

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 *LREC*, pages 1121–1126.