Eye Tracking as a Tool for Machine Translation Error Analysis

Sara Stymne, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull, Martin Wester

Linköping University, Sweden
Corresponding authors: {sara.stymne,henrik.danielsson}@liu.se

Abstract

We present a preliminary study where we use eye tracking as a complement to machine translation (MT) error analysis, the task of identifying and classifying MT errors. We performed a user study where subjects read short texts translated by three MT systems and one human translation, while we gathered eye tracking data. The subjects were also asked comprehension questions about the text, and were asked to estimate the text quality. We found that there are a longer gaze time and a higher number of fixations on MT errors, than on correct parts. There are also differences in the gaze time of different error types, with word order errors having the longest gaze time. We also found correlations between eye tracking data and human estimates of text quality. Overall our study shows that eye tracking can give complementary information to error analysis, such as aiding in ranking error types for seriousness.

Keywords: Machine translation evaluation, eye tracking, error analysis

1. Introduction

Evaluation of machine translation is a difficult task, both for humans and using automatic metrics. MT systems are often evaluated using automatic metrics, such as Bleu (Papineni et al., 2002), which commonly rely on comparing a translation to only a single human reference translation. Such a quantitative evaluation does not give any indications of the particular problems with a system. In addition, they need a large test set and the correlation with human judgments has been debated (Callison-Burch et al., 2006; Chiang et al., 2008). The most common type of human evaluation is estimations of adequacy and fluency, which can be useful, but is expensive and gives little information about particular strengths and weaknesses of the system. An alternative evaluation method is human error analysis, where errors in the MT output are identified and classified into different categories.

We have performed a preliminary study where we investigated the possibility of using eye tracking as a complement to other types of MT error analysis, by recording the eye movements of people reading machine translated texts. Our hypotheses were that bad MT output is harder to read than good MT output and that certain types of errors will take longer time for a reader to process.

2. Related Work

A very common way to evaluate MT systems is by using automatic metrics. The vast majority of automatic metrics, such as Bleu (Papineni et al., 2002) or Meteor (Denkowski and Lavie, 2010), are based on some way of calculating the closeness to one or more human reference translation, mostly giving a single system score for a collection of sentences. Using automatic metrics is fast and cheap, and can be useful, especially for comparing incremental versions of the same system, or for systems with a similar architecture. Metrics, however, usually only give a single quantitative score, and do not give much information about particular strengths and weaknesses of the system, even though different metrics focus on different aspects of the translation. Comparing scores from different metrics can give a very rough indication of major differences, especially in combination with a part-of-speech analysis (Popović et al., 2006).

Another evaluation possibility is human evaluation, which is often performed in order to compare several MT systems. It can be in the form of estimates of values such as adequacy and fluency, or by ranking sentences from different systems (e.g. Callison-Burch et al. (2007)). A combination of human and automatic metrics is human-targeted metrics such as HTER, where a human post-edits the output of a system to the closest correct translation, on which standard metrics such as TER is then computed (Snover et al., 2006). While both these types of evaluation are certainly useful, they are expensive and time-consuming, and still give only a quantitative score for each system, not telling us much about the particular errors of a system. These types of human evaluation work best with bilingual evaluators, who can compare the system output with the source sentence. It is also possible to present monolingual evaluators with one or several human reference translations as a source of comparison (Callison-Burch et al., 2007); this might, however, result in biased scores depending on certain choices made by the translator of the reference sentences.

An alternative type of human evaluation is error analysis, the identification and classification of MT errors. This type of evaluation is informative since it shows particular strengths and weaknesses for an MT system. It is, however, very time consuming to perform. There have been several suggestions for general MT error typologies that can be used for error analysis (Flanagan, 1994; Vilar et al., 2006; Farrús et al., 2010), targeted at different user groups and purposes. Flanagan (1994) also ranked error classes on two dimensions, improvability and intelligibility. There is no discussion of how this ranking was performed, however.

There have also been attempts of human evaluation without access to the source text or reference translation, such as evaluation based on reading comprehension or eye tracking. In reading comprehension studies, subjects read machine translated texts, and then answer reading comprehension questions about them (Fuji, 1999; Jones et al., 2005). Fuji (1999) found significant differences on reading com-

prehension questions, between texts with large quality differences.

Eye tracking is used to record a person's eye movement across a screen during tasks such as reading. From eye tracking equipment we can get measurements such as the count and duration of fixations, periods when our eyes remain relatively still. Humans can have more than one fixation on the same unit, so another common measurement is gaze time, the total time for all fixations on a unit. There have been numerous eye tracking studies of reading (see e.g. Rayner (1998) for a summary) but a general trend is that texts that are hard to read have more and longer fixations than easy texts. There is also a growing number of translation studies using eye tracking (e.g. Göpferich (2008); Pavlović and Jensen (2009)).

We are only aware of one study where eye tracking was used for MT evaluation (Doherty and O'Brien, 2009; Doherty and O'Brien, 2010). They performed a study where they investigated the use of eye tracking as a semi-automatic MT evaluation method. In their study they compared sentences that had been judged as excellent and poor in a previous human evaluation. They found that both average gaze time and fixation count was higher for the poor sentences than for the excellent sentences, but that there were no difference between the sentence sets on average fixation duration or pupil dilations.

Eye tracking studies where subjects are asked only to read the source documents can only be used to evaluate fluency, not adequacy, since a text can be well formed without reflecting the source document. Studies based on reading comprehension can be used for adequacy as well, if the comprehension questions cover relevant aspects of the source.

Our study differ from the study of Doherty and O'Brien (2010) in several ways. We let our subjects read coherent texts rather than isolated sentences. As Doherty and O'Brien (2010) point out, the reading patterns for single sentences has a reduced "ecological validity", since humans tend to read full texts rather than isolated sentences. We also analysed the eye tracking data on sub-sentential level, by looking at instances of errors, and do not only look at sentence level data. Doherty and O'Brien (2010) compare eye tracking measurements to HTER (Snover et al., 2006), adequacy, and fluency; whereas we use a human error analysis as the basis of our analysis, and also compare the eye tracking measurements to other data collected from the subjects in the study, such as reading comprehension questions and fluency judgements. They also picked out sentences from one MT system, which were ranked as either poor and excellent in a human evaluation, thus removing sentences with medium quality. We compare the output of three different MT systems, where two of them are of similar quality, while one is of a much lower quality. In both studies the focus in on fluency, not on adequacy, since only the source sentences are presented.

3. Experiment

We performed a user study where we recorded the eye movements of subjects when they read machine and human translated texts, which were translated from English to Swedish. We set up the experiment as a reading comprehension scenario, where the subjects were asked questions about translated texts after reading them. We recruited 33 university students as subjects for the user study. All were native speakers of Swedish except one, who had a very good command of Swedish. All subjects had a good command of English, which is an entry requirement to Swedish universities. Eleven of the subjects had to be dropped from the eye tracking analysis, since the eye tracking data for them were incomplete. The analysis of the other data is based on all 33 subjects.

The eye tracking was performed using SMI Remote Eye iView, an eye tracking system consisting of the eye tracking hardware and analysis software.[1] It is a non-invasive system, i.e., it does not require equipment like head-mounted displays or head-rests.

We based the analysis of the eye tracking data on areas-of-interest, boxes placed on the image used for eye tracking, that mark specific areas of the text. The measurements we were interested in, gaze time and number of fixations, are calculated for each box that marks an area-of-interest. Since we were interested mainly in error analysis we marked each error instance based on our human error analysis as an area-of-interest. Missing words were marked on the words surrounding the position where the missing word should have been. We will call such marking *error boxes*. As a point of comparison, we also marked correct words in the beginning, middle and end of each sentence, which we will call *control boxes*. When there was an error at a spot where we normally put a control box, we did not mark that spot, since we wanted control boxes only to cover fluent text. Gaze time and number of fixations were measured for error and control boxes, and in addition for the full texts.

3.1. Evaluation Specifications

We performed the evaluation on four short texts from Europarl (Koehn, 2005). The source texts had 504-636 words, enough to fill one screen in two columns in order to avoid scrolling. The average sentence length was 27 words for the English source, and 24 words for the Swedish reference translation. The texts were deliberately chosen to discuss four different subject matters: harbors, new EU members, renewable energy, and Russia, in order for the subjects not to be confused of the content in the different texts. All results were aggregated over the four texts per each system.

We manually performed an error analysis of the test texts from the three MT systems. The error analysis was performed with access to the English source. Errors were classified into the five base categories of Vilar et al. (2006): missing words, word order, incorrect words, unknown words, and punctuation; and as upper/lower case errors, which did not fit into the other categories. This is a relatively crude classification, and especially the incorrect category contains several types of errors such as agreement errors, extra words and incorrect word choice. Punctuation and upper/lower errors were ignored in the eye tracking analysis, since they were considered less prominent than

[1] See `http://www.smivision.com/`

the other error types, and since the errors made on these categories were quite similar for all systems. The error analysis was performed by two of the authors, both native Swedish speakers. On a sample analysis the two annotators had an error classification agreement of 87.8% (Kappa: 0.63).

3.2. MT Systems

We included three different English–Swedish MT systems in the study. All were standard phrase-based statistical machine translation systems, built using the Moses toolkit (Koehn et al., 2007) and trained on the Europarl corpus (Koehn, 2005). Two systems differ in the amount of training data: *Small* was trained on 100,000 sentences and *Large* on 701,157 sentences. The third system, *Comp* is trained with the same number of sentences as *Large*, but with the addition of a compound processing module (Stymne and Holmqvist, 2008). While the compounding module focused on processing compounds, this change also had other effects, since it affected the whole translation process, for instance by affecting the overall word alignment. We also compare the three MT systems to the human reference translation in Europarl, *Human*.

3.3. Procedure

Each subject read four different texts, one from each MT system and one human translation. The order and combinations of the texts and systems were balanced between the subjects. Each of the four texts was shown on the screen and the eye movements were recorded. The subjects were asked to read for comprehension and told that they would answer comprehension questions after they finished reading. There was no limit on the reading time; the subjects decided themselves when they had finished reading a text. After reading each text, the subjects were given a questionnaire with reading comprehension questions and estimation questions. There were three multiple-choice questions about the text content, and subjects were also asked to give confidence ratings of their answers on these questions. We also had three estimation questions, where subjects were asked to judge the fluency of the text, their experienced comprehension of the text, and the perceived amount of errors in the text on an 8-point scale.[2]

4. Results

In this section we first present the results on the contrastive evaluations: automatic metrics and error analysis. We then go on to discuss the results of the user study and the correlations between the different evaluation types. To calculate significance we used a repeated measures analysis of variance (ANOVA), except on the automatic metrics, where we used approximate randomization (Riezler and Maxwell, 2005).

4.1. Contrastive Evaluations

We evaluated the three SMT systems on two test sets, both the short texts used in the experiments, aggregated, with a total of 80 sentences and on a standard 2000 sentence Europarl test set. Table 1 shows Bleu (Papineni et al., 2002) and Meteor (Lavie and Agarwal, 2007) scores, on the two test sets, calculated based on one human reference. On both test sets, *Small* is significantly worse, on the 5%-level, than the other systems on both metrics. *Comp* is significantly better than *Large* on both metrics on the large test set, but on the short texts, there are no significant differences between these two systems, but the trend of which system is better is opposite on the two metrics. This contrast illustrates the challenge of evaluating systems with small quality differences on short texts.

	Short texts		Large test set	
	Bleu	**Meteor**	**Bleu**	**Meteor**
Comp	17.48	58.02	22.12	58.43
Large	16.96	58.58	21.63	57.86
Small	14.33	55.67	20.79	56.82

Table 1: Metric scores

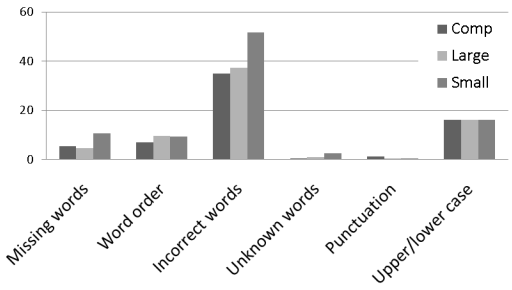

Figure 1: Frequencies of errors

Figure 1 shows the results of the error analysis, for the three MT systems. A repeated measures analysis of variance (ANOVA) showed significant differences between the three systems ($F(2,6) = 13.39, p < .05$), between the six error types ($F(5,15) = 41.84, p < .05$), and for the interaction between system and error type ($F(10,30) = 8.59, p < .05$).[3] The *Small* system has the highest number of errors, especially for incorrect and missing words, which is not surprising considering that it is trained on less data than the other systems, *Comp* has fewer errors than *Large* and incorrect words is by far the most common error in all systems.

4.2. User Study

For the full texts, there was no significant differences between all the translations either for overall gaze time or for fixation count. Errors had both a significantly higher number of fixations, 3.3 compared to 2.5 ($F(1,21) = 0.58, p < .05$) and a significantly higher average gaze time, 1418 ms compared to 998 ms ($F(1,21) = 8.55, p < .05$) than the control markers, also shown in Figure 2.

[2]The common feature *adequacy* could not be estimated, since the subjects did not see the source text.

[3]Standard notation for ANOVA results are used. In the formula $F(n,m) = x, p < .05$, F means that the F-test is used, n is the degrees of freedom for the between groups variance, m is the degrees of freedom for the error variance, x is the F-value and $p < .05$ means that the result is significant at the 5% level.

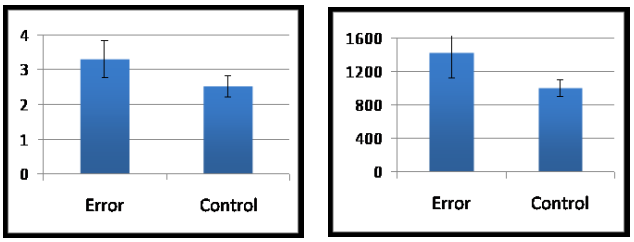

Figure 2: Average fixation count (left) and gaze time (right) for error and control boxes. Error bars show the 95% confidence interval.

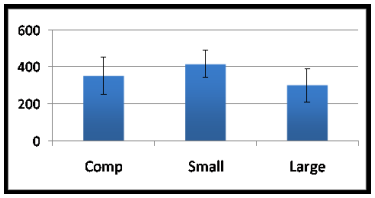

Figure 3: Average gaze time in the error boxes for the MT systems

There was also a significant difference between the average gaze time of errors in the three different systems ($F(1, 21) = 3.98, p < .05$), as shown in Figure 3. The *Small* system had longer average gaze time per error (415 ms) than *Comp* (350 ms) and *Large* (298 ms). This is different from the number of errors, which were fewest in *Comp* and is an indication that the errors that occur in *Small* and to some extent in *Comp*, might be more serious than the errors in *Large*, since they take longer time for readers to process.

The different types of errors have significantly different average gaze time ($F(3, 63) = 8.55, p < .05$), as shown in Figure 4. Word order errors have the longest average gaze time, followed by incorrect and missing words, with unknown words having the shortest time. All subjects had a good command of English; the fixations on the unknown English words would probably be more and longer with a source language that the subjects do not know.

The results on the reading comprehension and quality estimations are shown in Table 2. The differences between the four translations are not significant, but there are some overall trends. The number of correct answers on the reading comprehension questions is actually higher for the *Large* system than for the human reference, but the confidence of the correct answers is lower. On the estimation questions, the human translation is markedly better than all machine translated options. On both the estimation questions and reading comprehension, *Large* is best and *Small* is worse, with *Comp* in the middle.

We also investigated Pearson correlations between the eye tracking results and human estimates per system. For *Comp* there were significant correlations between total fixation count and estimated fluency, $r = -.39$ and estimated comprehension, $r = -.45$ and between total gaze time and estimated errors, $r = .37$ and estimated comprehension, $r = -.37$. For *Large* there were significant correlations between total fixation count and estimated fluency,

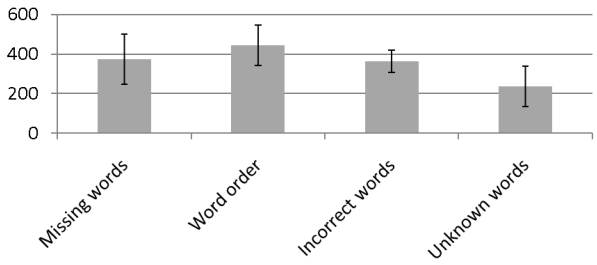

Figure 4: Average gaze time (ms) for different types of errors

$r = -.61$ and estimated errors, $r = .40$ and between total gaze time and estimated fluency, $r = -.38$ and estimated errors, $r = .37$. This shows that there are some moderate correlations between eye tracking measurements and human estimates, but they are not consistent for all systems.

5. Conclusion

We presented a preliminary study that showed that eye tracking can give information that complements other types of error analyses. Using Bleu or human estimates, it was hard to find differences between the systems, especially between the two best systems, *Comp* and *Large*. Using either error analysis or eye tracking, however, we were able to identify some differences between the systems.

We also showed that MT errors have both longer gaze times and more fixations than correct passages. Most importantly, we showed that the average gaze time is dependent on error types. This could be an indication that some error types are more disturbing for readers than others.

This study is small and preliminary, and there is plenty of room for more and larger studies on this theme. We would especially like to extend this study by using a more fine-grained error typology, since there likely are differences between the errors within each of our rather large error categories. It would also be interesting to test the methods on a post-editing scenario, on other language pairs, and on other translation systems. We also want to perform a qualitative investigation of parts in the texts that have long and many fixations. In our study we did not normalize for the size of the error boxes. While we strived to keep them of similar size by mainly marking one or two words, it would have been better to normalize the results based on box size.

We do, however, think there is a potential in eye tracking as a tool for error analysis. One clear possibility is to use eye tracking data to rank how serious different types of errors are, based on the number of fixations or gaze time of the error type. In this case it would also be possible to distinguish such rankings between different scenarios, such as reading MT output for comprehension, or post-editing it, by performing new eye tracking studies based on such scenarios. Another possibility could be to use eye tracking data on a text to mark places in the text that has long and many fixations, and thus are likely to be problematic in some way. Such markings could be useful for human error annotators. Another possibility is to try to predict error instances and types automatically based on eye tracking data.

	Correct answers	Confidence of correct answers	Estimated fluency	Estimated comprehension	Estimated errors
Human	64.50%	7.19	5.56	5.70	2.94
Comp	59.50%	6.43	3.50	4.85	5.67
Large	67.25%	6.82	4.16	4.86	5.34
Small	59.25%	5.97	3.33	4.53	6.11

Table 2: Results from questionnaire

6. Acknowledgements

We would like to thank Joel Johansson and Karin Ström Lehander for their help and guidance in using the eye tracking equipment. We also want to thank the anonymous reviewers for their useful comments.

7. References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *Proceedings of the 11th Conference of the EACL*, pages 249–256, Trento, Italy.

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*, pages 136–158, Prague, Czech Republic, June.

David Chiang, Steve DeNeefe, Yee Seng Chan, and Hwee Tou Ng. 2008. Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 610–619, Honolulu, Hawaii.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342, Uppsala, Sweden.

Stephen Doherty and Sharon O'Brien. 2009. Can MT output be evaluated through eye tracking? In *Proceedings of MT Summit XII*, pages 214–221, Ottawa, Ontario, Canada.

Stephen Doherty and Sharon O'Brien. 2010. Eye tracking as an MT evaluation technique. *Machine Translation*, 24(1):1–13.

Mireia Farrús, Marta R. Costa-jussà, José B. Mariño, and José A. R. Fonollosa. 2010. Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, pages 52–57, Saint Raphaël, France.

Mary Flanagan. 1994. Error classification for MT evaluation. In *Technology partnerships for crossing the language barrier: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, pages 65–72, Columbia, Maryland, USA.

Masaru Fuji. 1999. Evaluation experiment for reading comprehension of machine translation outputs. In *Proceedings of MT Summit VII*, pages 285–289, Singapore.

Susanne Göpferich. 2008. *Translationsprozessforschung: Stand, Methoden, Perspektiven*. Gunter Narr, Tübingen.

Douglas Jones, Edward Gibson, Wade Shen, Neil Granoien, Martha Herzog, Douglas Reynolds, and Clifford Weinstein. 2005. Measuring human readability of machine generated text: three case studies in speech recognition and machine translation. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1009–1012, Philadelphia, Pennsylvania, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL, Demonstration session*, pages 177–180, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, Pennsylvania, USA.

Nataša Pavlović and Kristian Jensen. 2009. Eye tracking translation directionality. In A. Pym and A. Perekrestenko, editors, *Translation Research Projects 2*, pages 93–109. Intercultural Studies Group, Tarragona.

Maja Popović, Adrià de Gisper, Deepa Gupta, Patrik Lambert, Hermann Ney, José Mariño, and Rafael Banchs. 2006. Morpho-syntactic information for automatic error analysis of statistical machine translation output. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 1–6, New York City, New York, USA.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at ACL'05*, pages 57–64, Ann Arbor, Michigan, USA.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human notation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Sara Stymne and Maria Holmqvist. 2008. Processing of Swedish compounds for phrase-based statistical machine translation. In *Proceedings of the 12th Annual Conference of the European Association for Machine Translation*, pages 180–189, Hamburg, Germany.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of machine translation output. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 697–702, Genoa, Italy.