Machine Translation Evaluation with Eye-tracking

Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz-Lillkull, Martin Wester, Linköping 2010.

Three variations of Machine Translation systems were evaluated. Four different texts have been machine translated from English to Swedish with the MT-systems. The translated texts have been evaluated and compared with a manually translated version generated from each original text. The quality has been assessed with error analysis and a reading comprehension test. An eye-tracking device has also been used for registering test subjects eye movement patterns when reading the texts. Eye movement data and answers to a questionnaire have confirmed that the systems are distinguished performance wise and that incorrect translations draw more attention. For one of the systems, there are correlations between how test subjects have ranked fluency, number of faulty translations and comprehension of the text, with the number of error fixations.

1 Introduction

Stymne and Holmqvist (2008) have studied how PBSMT can be improved by an algorithm that manages compound words in translation from English to Swedish. This system is called Comp and gives better translations when translated from English to Swedish than Large, without this algorithm according to automatic measurements done by Stymne and Holmqvist.

There are various methods for evaluating MT translations. One method for human evaluation described by Fuji (1999) is to let test subjects read the translation and afterwards answer to a questionnaire. The questions in Fuji's study concerned informativeness, comprehensiveness and fluency. Informativeness was measured with questions on the content, while comprehensiveness and fluency were measured on a four-point scale in which the test subjects were asked to assess how well they understood the text and how well they experienced fluency.

Doherty and O'Brien (2009) have done an assessment of eye-tracking as a methodology for evaluation of MT-systems. The variables being studied were average fixation time, duration per sentence/alphabet, comments of the test subjects and the correlation with earlier manual assessments. The results they received from the eye-tracking data correlates fairly well with the result from the manual evaluation.

Test subjects' rankings on fluency and correctness can be used as a measurement for manually evaluating reading comprehension (Gornostay, 2008). Based on Gornostay (2008) and Doherty and O'Brien (2009), this report describes an evaluation with eye-tracking measurements on Stymne and Holmqvist's (2008) systems.

1.1 Objective

The objective of this study is to evaluate different statistical models for machine translation that have been developed by Sara Stymne and Maria Holmqvist at the Department of computer and information science at Linköping University.

1.2 Question formulation

1. Does the type and number of faulty translations differ in the different MT-systems Comp, Large and Small?

2. If there is a difference between the systems regarding how errorful their outputs are, does the difference also ascribe to measured fluency?

3. Are there any correlations between the measured fluency in the eye-tracker and the estimated fluency, comprehension and quantity of faulty translations in the questionnaire?

1.3 Hypothesis

1. We are going to see a difference in the translation quality among the MT-systems, in accordance with the automatic evaluation done by Stymne and Holmqvist.

2. We are going to see a difference in fluency when reading different translations generated by the different variations of MT-systems. According to the eye-tracking data, faulty translations should draw more fixations and have a longer fixation time than the correct translations.

3. There is a correlation between fluency measured by the eye-tracker and what the test subjects have experienced in terms of fluency, quantity of faulty translations and comprehension.

2 Methodology

A combination of error analysis, a reading comprehension questionnaire and eye-tracking measurement have been used in this study to evaluate different variations of three MT systems. The error analysis has been implemented in the eye-tracker analysis program to show how the different systems perform. The eye-tracker has also been used for collecting reading behaviors and questionnaires have been used for capturing test subjects' impressions of the texts.

2.1 Test subjects

The subjects were 33 students at Linköping University aged from 20 to 30 years old, 22 men and 11 women.

2.2 Texts

In the tests, 4 texts from the European parliament from year 1999 have been used. They were machine translated from English to Swedish with the systems that were being evaluated. A manually translated text, called gold standard, was added as a control text.

2.3 Questionnaires

Four different questionnaires were created for the four text domains. There were six questions in total per questionnaire. Three of them were related to the text, also called comprehension questions, and the other three were related to the reading experience, also called evaluation questions.

2.4 MT systems

The systems Large and Comp, acquainted from Stymne and Holmqvist (2008) were used in the study. Large is a standard statistic translation system, and Comp differs from Large by the added algorithm that manages compound words. Small has also been used for translations. It uses a corpus with only 100 000 sentences as training data, compared to more than the 700 000 sentences used in the other systems. The manually translated text was used as a golden standard to be compared with the other translations.

2.4 Procedure

The evaluation was made on four different text domains each translated by Large, Small and Comp, with an additional gold standard. These 16 translations were divided into eight different tests. Each test consisted of four texts, one text from each text domain. Three of these texts were translated with different MT systems and one was the gold standard.

The system which has been used in this study is iView and it was developed by the company SMI (SensoMotoric Instruments). The eye-tracking system can among other things capture the time the

eye fixates at one particular stimuli and how many fixations that are made (SMI, 2009). The test subjets were instructed to sit down in front of the computer screeen which had the eye-tracker camera attached to it underneath. The stimuli pictures were suited to fit the four texts. The subjects were handed a questionnaire after reading each text, and before the test they were told that the questions would relate to the content of each text. The point of giving these instructions was to increase the test subjects motivation to understand the text. There was no time limit during the test.

2.5 Analysing methods

The translated texts, the questionnaires and the data from the eye-tracker all required different methods for their analyses.

2.5.1 Error analyses

The twelve machine translated texts were error analysed manually. The errors generated at translation were divided into six categories, based on Fuji (1999). These categories were missing words, word order, incorrect words, unknown words, punctuation and upper/lower case letters.

2.5.2 Questionnaires

The three initial questions of the questionnaire were corrected. Correlation was calculated between the results of the questionnaire and data from the eye-tracker.

2.5.3 Eye-tracker

The collected data was analysed by marking the areas containing errors found in the error analyses. All words that were part of an error category were Comp with an AOI (Area of Interest) that was numbered depending on error type. To be able to compare the values of the error AOIs, control AOIs were also placed in different places in the text. They were placed whenever a correct translation occurred in the beginning, the middle and the end of a sentence. The control AOIs were not supposed to overlap the error AOIs, so when there was an error AOI in the beginning, middle or end of a sentence, no control AOI was placed there.

4 Results

The results from the different parts of the study are presented as follow. All the declared results are significant if nothing else is mentioned. Only the most valuable results are accounted for.

4.1 Error analysis

When comparing the systems based on the number of errors they generated we found that the three systems Large, Comp and Small were different, F (2,6) = 13.39, p < .05, η^2 = .82. We also found that how common the various kinds of errors were differs, F (5, 15) = 41.84, p < .05, η^2 = .93. The interaction between translation and error type was significant, F (10, 30) = 8,59, p < .05, η^2 = .74. For these results, see Figure 1.

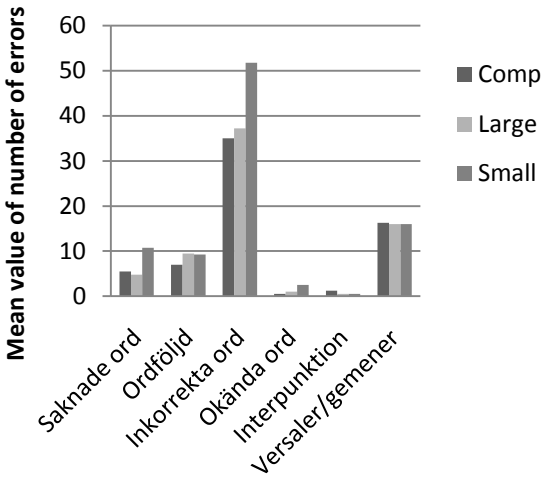

Figure 1. Frequency of errors for each error type in the three MT systems.

4.2 Questionnaire

Question one to three consisted of questions about the content, no significant difference of the answers was to be found between the systems. Question four to six concerned how the text was experienced

by the test subjects. There were on estimated fluency, number of errors in the text and comprehension of the text no significant difference between the translations, though for all the systems we found a strong positive correlation between estimated fluency and estimated comprehension, as shown below.

Comp: ρ(estimated fluency, estimated comprehension) r = .58

Large: ρ(estimated fluency, estimated comprehension) r = .63

Large little ρ(estimated fluency, estimated comprehension) r = .59

4.3 Eye-tracking

The boxes containing errors have a higher fixation time in total than the control boxes, F (1, 21) = 8.55, $p < .05$, $\eta^2 = .29$, see Figure 2. The error boxes also have a larger number of fixations than the control boxes, F (1, 21) = .58, $p < .05$, $\eta^2 = .03$, see Figure 3.

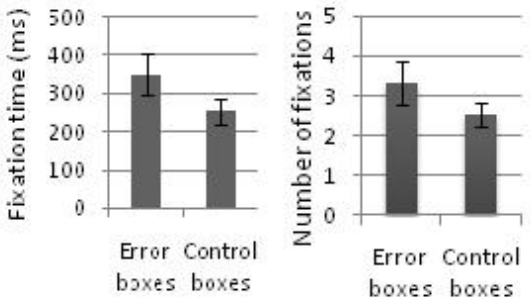

Figure 2 and 3. Mean of fixation time and number of fixations for the error boxes' AOI and the control boxes' AOI.

Regarding length of fixation a strong significant difference was to be found between the translations from the the gold standard and Small, F (1, 21) = 3.45, $p < .05$, $\eta^2 = .14$, see Figure 4.

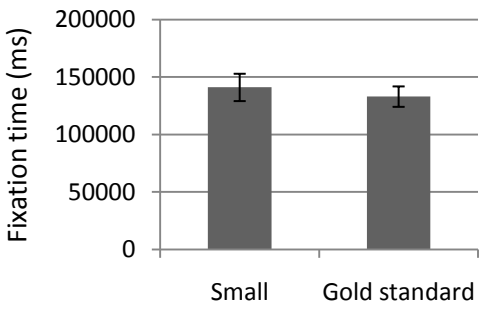

Figure 4. Difference in fixation time between Small and the gold standard, for the entire texts.

No significant difference was found between the other systems.

A significant difference was also found concerning mean fixation time for the error boxes, F (2, 42) = 3.98, $p < .05$, $\eta^2 = .16$, and for the control boxes, F (2, 42) = 5.44, $p < .05$, $\eta^2 = .21$, see Figure 5.

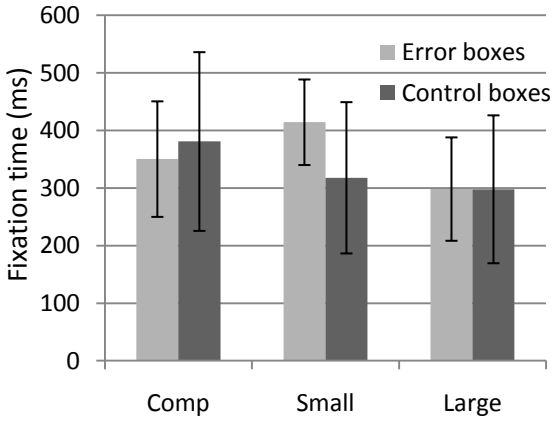

Figure 5. Difference in mean fixation time for the error boxes in the translations generated by the three systems.

4.4 Correlation between eye-tracking data and questionnaire

A strong negative significant correlation we found between the eye-tracking data and the questionnaire was in Large between estimated fluency and number of fixations.

ρ(fluency, fixation count) r = -.61

Another weak negative significant correlation found in Comp was between estimated comprehension and total number of fixations.

ρ(comprehension, fixation count) r = -.45

5 Result discussion

The first hypothesis was that the systems would differ in terms of number and type of faults. The significant results from the error analysis favours this hypothesis when it comes to how many errors the systems generate, and that Small performs slightly poorer than the other two MT systems.

The boxes containing errors have longer fixation time than the control boxes and they also have a larger number of fixations. Our study showed that mistranslated or missing words effect the total fixation time and number of fixations. We think the explanation for this result is that sentences containing errors give rise to more attention than sentences with correct semantic and syntactic content.

Two of the MT systems' translations did neither differ concerning length of fixation time nor number of fixations. This contradicts the second hypothesis that there is a relationship between how the systems perform due to the analysis of errors and how they perform due to the eye-tracking measurements. Though, we could see a strong difference between the gold standard and Small, which, in spite of the other results, indicates that there might be a small difference between the systems. It might be harder to detect such a difference via eye-tracking than with other methods, which would explain why our results from the eye-tracker are not very clear.

When comparing the eye-tracking data with the answers to the questionnaires we found that in Large fewer fixations correlated with better estimated fluency, and in Comp fewer fixations correlated with better estimated comprehension. We believe that the reason for this is that fewer fixations infer higher fluency, and that the subjects also perceive the text to have higher fluency, which contribute to a better understanding of the whole text. These results favours our third hypothesis, though we could not see the effect in all the systems.

6 Method discussion

Words that have been tagged as belonging to an error category do not necessarily have to be incorrect in either a grammatical or semantic way. One example of a Comp area where the words were incorrectly translated is "jag skulle vilja påpeka att jag ser förmiddagens debatt **inom ramen för** *(relatera till)* europeiska unionens förmåga att reformera.". The sentence can be perceived as correct with the incorrect words.

The eye-tracker is relatively sensitive concerning disturbance. If the camera does not succeed in finding the subject's eyes, the system cannot record any data. This problem occurred during some of the tests, which resulted in data from these subjects was recorded only during the time their eyes were localized by the eye-tracker. Thanks to numerous participants, this should not have affected our results in a big way.

7 Conclusions and future work

The results from the error analysis showed that there is a significant difference among Stymne and Holmqvist's systems regarding which type of errors and numbers of errors the systems generates. Significant results were also found between how often the different types of errors occurred in the translations and that errors draw more and longer fixations. However, no difference between the four translations was to be found was to be found when it comes to comparing number and length of fixations for the entire texts.

In the questionnaire there was a strong correlation between estimated fluency and estimated comprehension. We think this is because whether the comprehension is good or bad for a text depends on how good the fluency is.

The results concerning the quality of the translations did not agree very well between eye-tracker and other measurements. The hypothesis was that translations containing more errors would attract more and longer fixations. Since no difference was to be seen between the system, we think it should be further investigated if eye-tracking really is a suitable method for evaluation of different MT systems, or whether another approach is preferable.

References

Doherty, S. and O'Brien, S. (2009). Can MT output be evaluated through eye tracking? In *MT summit*, pages 214-221. Ottawa, Canada.

Fuji, M. (1999). Evaluation Experiment for Reading Comprehension of Machine Translation Outputs. In *MT Summit VII*, pages 285-289. Singapore.

Gornostay, T. (2008). Machine translation Evaluation. Unpublished course report, *NGSLT Machine Translation course*.

SMI (SensoMotoric Instruments) (2009). *iView X System Manual. 2.2 ed.*

Stymne, S. and Holmqvist, M. (2008). Processing of Swedish Compounds for Phrase-Based Statistical Machine Translation. In *Proceedings of EAMT08, European Machine Translation Conference*, pages 180-189. Hamburg, Germany.