A User-Based Usability Assessment of Raw Machine Translated Technical Instructions

Stephen Doherty
Centre for Next Generation Localisation
Centre for Translation & Textual Studies
Dublin City University
stephen.doherty@dcu.ie

Sharon O'Brien
Centre for Next Generation Localisation
Centre for Translation & Textual Studies
Dublin City University
sharon.obrien@dcu.ie

Abstract

This paper reports on a project whose aims are to investigate the usability of raw machine translated technical support documentation for a commercial online file storage service. Following the ISO/TR 16982 definition of usability - goal completion, satisfaction, effectiveness, and efficiency - comparisons are drawn for all measures between the original user documentation written in English for a well-known online file storage service and raw machine translated output in four target languages: Spanish, French, German and Japanese. Using native speakers for each language, we found significant differences ($p < .05$) between the source and MT output for three out of the four measures: goal completion, efficiency and user satisfaction. This leads to a tentative conclusion that there is a difference in usability between well-formed content and raw machine translated content, and we suggest avenues for further work.

1 Introduction

It is generally agreed that, in commercial contexts, machine translation (MT) output still needs to be post-edited in order to be acceptable to, and usable by, end-users. However, a number of factors have generated interest in the possibility of using raw, i.e. non post-edited, MT output. First, the quality of raw MT output has improved significantly over recent years, thanks to improvements in NLP techniques in general. This is true for some languages, but not for all. Second, the pervasiveness of online MT systems has resulted in users making use of raw MT output for their own purposes, usually defined in MT circles as 'gisting'. Third, commercial companies who want to implement MT as a way of dealing with increasing volumes and pressure to decrease costs have met with significant opposition from their translation supply base and this has forced a discussion on how to by-pass post-editing or find alternative resources (e.g. domain experts or crowd-sourced volunteers) to post-edit. This situation forces the question: just how usable is raw MT output?

There are relatively few studies of the *usability* of raw machine translated documentation by real end-users. For example, Tomita's work (Tomita, 1992; Tomita *et al.*, 1993) focused on the concept of content comprehension. Fuji (1999) evaluated the *informativeness*, *comprehension*, and *fluency* of MT output, where participants had no reference to the source text, while Fuji *et al*. (2001) measured the concept of *usefulness*. Jones *et al.* (2005) measure the *readability* of MT output. While comprehensibility and readability are frequently considered to be components of usability, these studies address only specific aspects of the concept of usability. Gaspari's study

(2004) in which real end users' needs are evaluated in the context of web usability comes closest to the study presented here. However, Gaspari's focus is on the usability of online MT systems, as opposed to the text they generate.

This paper reports on a project whose aims are to investigate the usability of raw machine translated technical support documentation for a commercial online service. It builds on previous work which investigates the use of eye tracking as a machine translation evaluation mechanism (Doherty and O'Brien, 2009; Doherty *et al.*, 2010, which focused on the readability and comprehension of machine-translated technical support documentation (Doherty, 2012), and on the impact of controlled authoring on the readability of MT output (O'Brien, 2010).

While there is some divergence around the definition of usability, the majority of terms in the literature closely adhere to the ISO definition. Following the ISO/TR 16982 definition, usability is understood here as "the extent to which a product can be used by specified users to achieve **specified goals** with **effectiveness**, **efficiency**, and **satisfaction** in a specified context of use" (ISO, 2002). The objective of this study was to establish how usable raw machine translated instructions were for end users in comparison with the original source text, which was in English.

2 Method

To ensure the task was as realistic as possible, we selected English documentation for a well-known online file storage and sharing service. We made an initial assumption that the original English instructions published by the developer were reasonably usable, given that the service has over 50 million users (Barret, 2011). As native speakers of English, both authors judged the documentation to be of reasonable quality and well-formed. These were initial assumptions which would be tested in the project.

Source documentation was selected from the service's support database and modified to produce six sequential tasks to provide a realistic first session for the user. The authors wanted a series of coherent instructions so that participants could be tested on task completion and efficiency (described later in this section).

A non-domain specific freely available machine translation system was used to translate the documentation. This system was selected also because the scenario of a real end user using this type of system (as opposed to a domain-specific or in-house engine) to translate documentation for comprehension purposes was realistic. The documentation was translated into French, German, Spanish, and Japanese as these were the languages for which we could recruit native speakers as experimental participants and for which the software developer provides a partially localized interface. Twenty-nine participants were recruited (English = 15, French = 4, German = 3, Spanish = 3, Japanese = 4). The criteria for inclusion as participants were: (1) the participant was a native speaker of the target language; (2) they had not yet signed up to or used the online service but (3) they were a prospective real user because they use computers and create electronic documents on a daily basis and could, therefore, potentially avail of the online service to store and/or share files, and (4) they were willing to give consent to participate in a research project involving eye tracking and other measurements.

For the purposes of this paper, all native speakers of English are placed in one group (the non-MT group, n = 15) and all target language participants are placed in one group (the MT group, n = 14). This enables comparisons between the MT and non-MT groups, which are of almost equal number. Comparisons between the non-MT group and each specific target language will be reported at a later date.

In keeping with the ISO definition given above, the measures of usability in this study are:

1. Goal completion: a measurement of the success or failure of the tasks, which were guided by the documented instructions;
2. Total task time (effectiveness): a measurement of the overall duration of the tasks in seconds;
3. Efficiency: measured as the number of successful tasks completed (out of all possible tasks) when total task time is taken into account;

4. Satisfaction: a measurement of user satisfaction of the instructions on a post-task 5-point Likert scale.

A Tobii 1750 eye tracker was used to record the reading of the instructions, placed on the left-hand side of the screen and the task execution, which occurred on the right-hand side of the screen (see Figure 1). The eye tracker captured measures 1-3, while the post-task questionnaire captured measure 4.

Figure 1. Screen Setup

3 Results

3.1 Goal Completion

Each task was assigned one or zero points depending on whether the task could be completed fully, e.g. deleting a file, or not at all – one point was earned for each successful task with a maximum of 30 points in total. An independent samples t-test found a significant difference between the conditions (t = 2.312, df = 14.312, p = .036) where the source condition resulted in an average score of 29.73, (median = 30, SD = .594, min. = 28, max. = 30), compared to the target where the mean was 28 (median = 29.50, SD = 2.746, min. = 22, max. = 30) – see Figure 2.

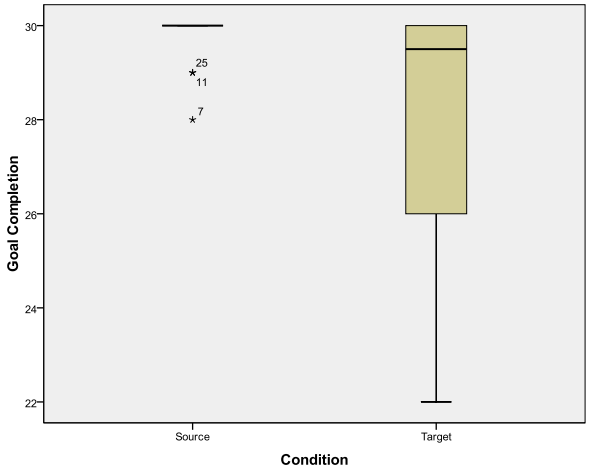

Figure 2. Goal Completion for Both Conditions

3.2 Total Task Time (in seconds)

An independent samples t-test found no significant difference between the conditions (t = -1.177, df = 26, p = .25) where the source text resulted in a lower mean of 437.30 seconds (median = 321.40, SD = 308.94, min. = 165.80, max. = 1352.70) than the target, whose mean was 588.85 seconds (median = 391.80, SD = 372.32, min. = 209.00, max. = 1267.40) – see Figure 3.

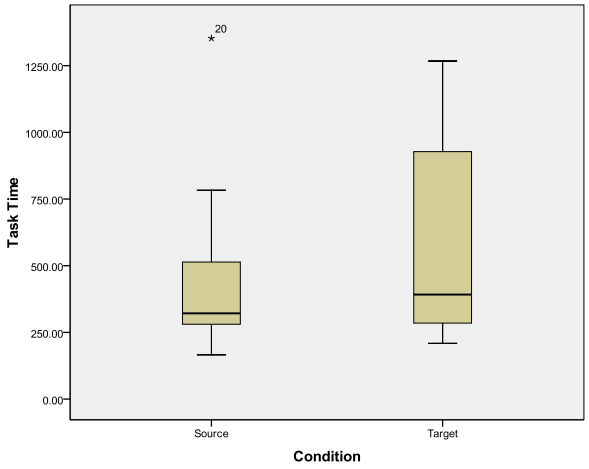

Figure 3. Total Task Time in Seconds for Conditions

3.3 Efficiency

As described above, efficiency is calculated by the number of successful tasks

completed against the total number of tasks in the experiment, then expressing this result as a divisor of the respective total task time:

$$\sum \frac{efficiency}{total\_task\_time(sec.)} x100, where \frac{task\_sucesses}{total\_tasks} x100 = efficiency$$

For example, participant 1 completed all thirty tasks in 579.40 seconds, giving an efficiency score of 17 compared to participant 2 who also complete all thirty tasks but took 1185.40 seconds, resulting in a score of 8 – in this way we report that participant 1 is more efficient, and has a higher value for this measure.

An independent samples t-test found a significant difference between the conditions (t = 2.085, df = 27, p = .047) where the source condition was more efficient (mean = 31, median = 31, SD = 15.6, min. = 7, max. = 60) than the target (mean = 20, median = 19, SD = 11.4, min. = 7, max. = 38) – see Figure 4.

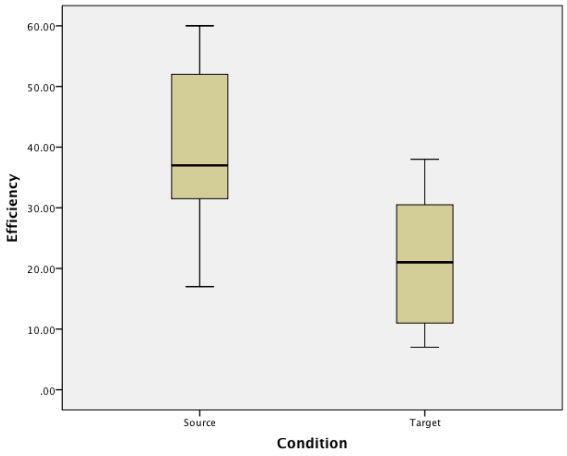

Figure 4. Efficiency for Both Conditions

3.4 Satisfaction

As measured via the post-task questionnaire on a 5-point Likert scale, users were asked to indicate to what extent they agreed with the statement: "I was satisfied with the instructions provided", where 1 meant *strongly disagree*, and 5 meant *strongly agree*. An independent samples t-test found a significant difference between conditions (t = 3.373, df = 20.271, p = .003) where the source text received higher levels of satisfaction with a higher mean of 4.13 (median = 4.0, SD = .743, min. = 2, max. = 5) to the target condition's 2.79 (median 2.5, SD = 1.311, min. = 1, max. = 5) – see Figure 5.

3.5 Correlational Analysis

Table 1 shows the correlation coefficients for each of the above variables. As indicated in Figure 6, efficiency is strongly correlated with task time (ρ = -.978, p < 0.01) and moderately correlated with satisfaction (ρ = .442, p < 0.05), however, this is not the case for goal completion (ρ = .185, p > 0.05) due to the majority of participants completing all of the tasks successfully.

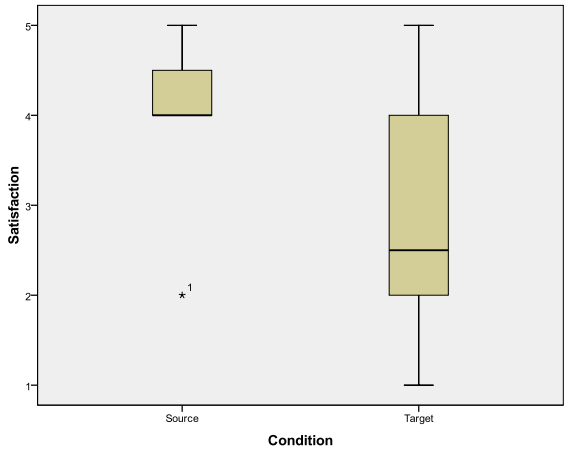

Figure 5. User Satisfaction for Both Conditions

Goal Completion	-	.63**	-.34	.47**
Satisfaction	.63**	-	.55**	.52**
Task Time	-.34	-.55**	-	.75**
Efficiency	.47**	.52**	-.75**	-

Table 1. Correlation Coefficients (ρ) for Each Variable[1]

[1] ** Correlation is significant at the .01 level (2-tailed).

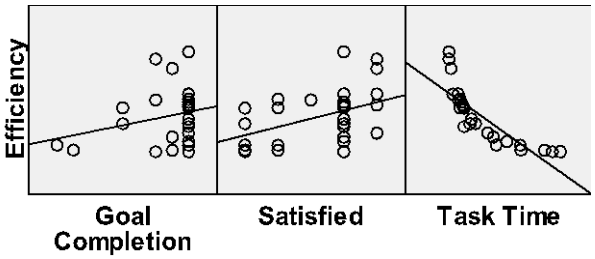

Figure 6. Correlation Coefficients (ρ) for Efficiency

4 Summary & Conclusion

Our aim was to investigate the usability of raw machine-translated instructional content for a variety of target languages and to compare that to the usability of the source English content using eye tracking and screen recording to capture user interaction and applying four measurements based on the ISO/TR 16982 definition of usability: goal completion, total task time, efficiency and user satisfaction ratings. By utilizing technical support content from a popular online storage service, and native speakers of each of the above languages, we created a strongly ecologically valid scenario.

We found significant differences ($p < .05$) between the source and MT output for three out of the four measures: goal completion, efficiency and user satisfaction. This leads us to a tentative conclusion that there is a difference in usability between well-formed content and raw machine translated content.

This phase of the research divided participants into 'source' and 'machine translated' groups, but there were four distinct languages within the machine-translated group. The next phase will involve looking at similarities and differences across the four target languages.

We have also collated eye tracking measurements such as total fixation counts, average fixation duration and percentage change in pupil dilation, all of which are shown to be indicators of cognitive load (Duchowski, 2007).

Another interesting question is whether the usability measurements will differ if (1) the raw MT content is post-edited and (2) if it is translated by human translators. It is our hope to make comparisons between these different content types in the future.

Acknowledgments

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University.

References

Barret, V. 2011. Dropbox: The Inside Story of Tech's Hottest Startup". *Forbes Magazine*, Nov., 7th. Available from: http://www.forbes.com/sites/victoriabarret/2011/10/18/dropbox-the-inside-story-of-techs-hottest-startup/print/

Doherty, S. & O'Brien, S. 2009. Can MT output be evaluated through eye tracking? *MT Summit XII: Proceedings of the Twelfth Machine Translation Summit*, Ottawa, Ontario, Canada, pp. 214-221.

Doherty, S., O'Brien, S. & Carl, M. 2010. Eye tracking as an MT evaluation technique. *Machine Translation*, **24**(1). Springer.

Doherty, S. 2012. *Investigating the effects of controlled language on the reading and comprehension of machine translated texts: A mixed-methods approach.* PhD thesis, Dublin City University. [Available from: http://doras.dcu.ie/16805/]

Duchowski, A. T. 2007. *Eye Tracking Methodology: Theory and Practice*. Springer-Verlag, London.

Fuji, M. 1999. Evaluation experiment for reading comprehension of machine translation outputs. *Proceedings of MT Summit VII*, pp. 285-289.

Fuji, M., Hatanaka, E., Ito, S., Kamai, H. Sukehiro, T., Yoshimi, T., & Ishara, H. 2001. Evaluation method for determining groups of users who find MT useful. Proceedings of *MT Summit VIII: Machine Translation in the Information Age*, Santiago de Compostela, Spain, 18-22 September, pp.103-108.

Gaspari, F. 2004. Online MT Services and Real Users' Needs: An Empirical Usability Evaluation. In Frederking, R. E. & Taylor, K. B. (eds.), *Association for Machine Translation in the Americas 2004*, LNAI 3265, pp. 74–85.

International Organization for Standardization. 2002. ISO/TR 16982: Ergonomics of human-system interaction – Usability methods supporting human-centred design.

Jones, D., Gibson, E., Shen, W., Granoien, N., Herzog, M., Reynolds, D., & Weinstein, C. 2005. Measuring human readability of machine generated text: three case studies in speech

recognition and machine translation. *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, pp.1009-1012.

O'Brien, S. 2010. Controlled language and readability. In Shreve, G. M. and Angelone, E. (eds.), *Translation and Cognition*. American Translators Association Monograph Series XV. Philadelphia: John Benjamins.

Tomita, M. 1992. Application of the TOEFL Test to the Evaluation of Japanese-English MT. *Proceedings of MT Evaluation Workshop, AAMT*, Nov. 1992.

Tomita, M., Shirai, M., Tsutsumi, J., Matsumura, M. & Yoshikawa, Y. 1993. Evaluation of MT Systems by TOEFL. *Proceedings of the Theoretical and Methodological Implications of Machine Translation*, TMI-93.