

Automated Machine Translation Improvement Through Post-Editing Techniques: Analyst and Translator Experiments

Jennifer Doyon

Abraxas Corporation
Herndon, VA 20170 USA
jen.doyon@abraxascorp.com

Christine Doran

C. Donald Means
The MITRE Corporation
Bedford, MA 01730 USA
cdoran@mitre.org
donmeans@mitre.org

Domenique Parr

The MITRE Corporation
McLean, VA 22102 USA
dtp23@georgetown.edu

Abstract

From the Automatic Language Processing Advisory Committee (ALPAC) (Pierce et al., 1966) machine translation (MT) evaluations of the '60s to the Defense Advanced Research Projects Agency (DARPA) Global Autonomous Language Exploitation (GALE) (Olive, 2008) and National Institute of Standards and Technology (NIST) (NIST, 2008) MT evaluations of today, the U.S. Government has been instrumental in establishing measurements and baselines for the state-of-the-art in MT engines. In the same vein, the Automated Machine Translation Improvement Through Post-Editing Techniques (PEMT) project sought to establish a baseline of MT engines based on the perceptions of potential users. In contrast to these previous evaluations, the PEMT project's experiments also determined the minimal quality level output needed to achieve before users found the output acceptable. Based on these findings, the PEMT team investigated using post-editing techniques to achieve this level. This paper will present experiments in which analysts and translators were asked to evaluate MT output processed with varying post-editing techniques. The results show at what level the analysts and translators find MT useful and are willing to work with it. We also establish a ranking of the types of post-edits necessary to elevate MT output to the minimal acceptance level.

1 Introduction

With research and development projects in machine translation dating back to the early 1950s, several approaches have been invented and applied to the creation of MT engines over the past 50+ years.

The best known MT strategies include: Direct, Transfer, Interlingua, Example-based, Statistical, and most recently Hybrid, the combination of both Transfer and Statistical approaches. Regardless of the approach to machine translation employed, MT output quality has historically been considered "poor." Recognizing that the definition of "poor" varies by system and by intended use of MT output, the PEMT project sought to provide a more explicit definition of "poor" MT output and to explore the ability to improve the quality of machine translation output through a number of post-editing strategies.

2 Overview

The three main activities of the PEMT project were identifying and categorizing MT output errors; applying post-editing techniques to the MT output; and evaluating and analyzing the resulting effect on the quality of the MT output in order to establish a user acceptance level. Throughout this project, the team focused its efforts on correcting fluency errors (i.e., errors affecting the intelligibility or grammaticality of MT output), which tend to impact user acceptance of MT the most. In our first year, we focused on a single Transfer-based, Arabic-to-English MT system. In the second year, we added two more Transfer-based engines, still for Arabic-to-English. All three of these MT engines have been and continue to be funded and/or utilized by the U.S. Government.

During the first year of this project, the PEMT team explored the use of second language learning commercial-off-the-shelf (COTS) products to perform error diagnosis and remediation of raw MT output. Unfortunately, only one of the ten commercial post-editing tools evaluated yielded any statistically significant improvement in MT quality.

In Year 2, we expanded upon this finding while exploring novel MT post-editing methods, by applying refined COTS post-editor feature settings and then expanding into the domain of human post-editing in an effort to evolve beyond the state-of-the-practice in MT post-editing techniques. We performed experiments with both analysts and translators to determine progress toward obtaining and/or surpassing the user-determined MT acceptability level. The proceeding sections of this paper will discuss the Analyst and Translator Experiments performed and Results reported in Year 2 of the PEMT project.

3 Analysts and Translator Experiments

The PEMT project conducted two sets of experiments, one with analysts and one with translators. Both conditions used web-based survey software and collected basic background and demographic data and comments in addition to the evaluators' judgments. In the Analyst Experiment, analysts were shown 30 documents and asked to rate the grammaticality of a given version of an English translation on a seven-point scale, in answer to the question "How acceptable is the grammar in this passage?" The questionnaires were comprised of five different randomized sets of 30 documents, such that any given participant would score each tool twice (two different outputs from each of the 15 versions of post-edited data, including unedited MT). Each of the five sets paired different documents with the tools (for a total of ten different documents per tool) to ensure any effects found would not be due to a particular document. Participants were randomly distributed to one of the five questionnaires. In addition, the 30 documents were randomized within each questionnaire.

In addition to rating the grammaticality of the passages, the analysts were also asked to indicate at what level on the scale a document became useful to them. Our goal was to identify the grammatical features of a translation that affected analysts' perception of its usability.

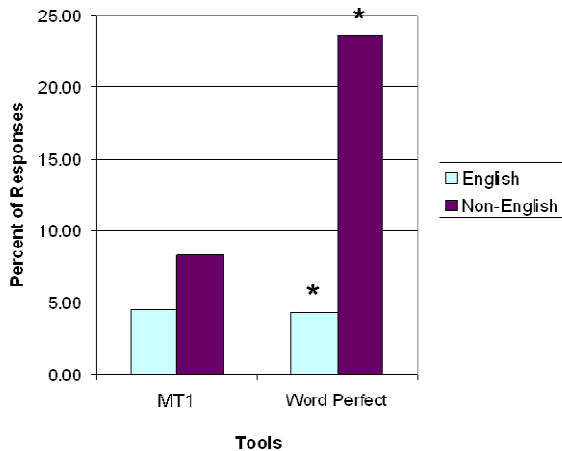
The minimum number of analysts needed to participate in this experiment to produce statistically powerful scores was 30, with any additional participants increasing the sensitivity of the measurements. This number allowed for approximately six participants to score each of the five versions of the evaluation. Because earlier results from Year 1

showed no significant judgment differences between analyst and non-analysts, the participant pool included engineers, administrative assistants, managers, and other professionals. The recruitment process resulted in the participation of 6 analysts and 31 non-analysts.

In the Translator Experiment, Arabic-to-English human translators were presented with an Arabic source text and one of its corresponding English target translations from the experiment corpus and asked to judge the translations based on three levels of usability. The question they were asked was "If you had to translate this document, would you prefer to edit the English target passage, translate the Arabic using the English passage as a reference, or translate the Arabic from scratch without referring to the translated passage?" As in the Analyst Experiment, our goal was to determine the grammatical features of a translation that affected translators' perception of its acceptability.

The ideal number of translators needed to participate in this experiment to produce statistically powerful scores was 20. This number allowed for four translators to score each of the five versions of the evaluation. The recruitment process yielded 21 Arabic-to-English translators. Five characterized Arabic as their native language, the remaining 16 were native English speakers. All Arabic native speakers were near-native English speakers.

In addition to performing the revised Translator Experiment in Year 2, we also conducted a second version of the Translator Experiment from Year 1, which had included mostly non-native, non-near-native English speakers, in order to determine whether there were differences between the responses of native or near-native and non-native English speakers. The results from this experiment can be seen in Figure 1.



* Indicates a significant difference between the ratings for English and Non-English speakers for Edit response.

Figure 1. Year 1 Native English Speakers vs. Year 1 Non-Native English Speakers

While the responses did not significantly differ for the raw output of system MT1, they did significantly differ for MT1 post-edited by WordPerfect, with native English speakers being less likely than non-native speakers to edit documents post-edited by WordPerfect. This result could be attributed to the fact that native and near-native English speaking translators are more discriminating (i.e., more critical of small grammatical differences than non-native translators.)

3.1 Experiment Corpus

Our corpus was taken from a collection of 81 Arabic broadcast news and newswire texts translated into English, averaging 300-500 words in length. For Year 1, 36 documents were selected from this larger collection to be translated into English using MT1's engine, cropped to 100-150 words (to avoid a bias against longer documents, which naturally contain more errors), and then passed through ten commercially available post-editing tools using their default settings. Seven of these ten tools were also run with the spell checking feature turned off because in many cases the spell checking feature was actually introducing errors. In the remaining three tools, spell checking could not be disabled or was not part of the tool. This gave us a total of 18 versions of each text, including the raw MT output.

For the Year 2 experiments, 30 different documents were selected from the collection of 81 Arab-

ic texts. These texts were translated into English using the MT1, MT2, and MT3 engines. The English output from each of these three systems was then post-edited using two human and two automated methods, each of which will be discussed in detail below. This resulted in a total of 15 versions of each document, including the raw MT output. Again, all texts were cropped before being presented to the analysts and translators.

3.2 Human Post-Edits

The goal here was twofold: 1) to see what could be learned from human editors, with the intention of analyzing and incorporating their more amenable techniques into an automated post-editing tool and 2) to explore the feasibility and value of using human-edited MT output. We used two types of human edits: Full Edits and Brief Edits.

For the Full Edits, the English MT output was split between three professional editors, such that each worked on documents from each of the three engines. These native English-speaking editors, who had no knowledge of Arabic, were instructed to produce publication quality edits, as opposed to performing syntax-only corrections.¹ They worked only on the raw English translations. Fifteen edited documents (five from each editor) were then selected and analyzed by one of the PEMT team's linguists both for descriptive purposes and for use in the development of an auto-correction tool. A total of 588 edits were identified, of which 60% were determined to be beneficial (e.g., changing word order) and the remaining 40% were determined to be neutral, or non-detrimental, primarily changing/deleting one or more English words. The majority category was change/delete English word(s) (e.g., *external* to *foreign*) encompassing 37% of the total errors identified in the Full Edits. In analyzing the Full Edits, we found that the editors were rarely able to handle transliterated Arabic text and that the changes produced by the Full Edits were much more style-based than content-based (no surprise given that they did not have the Arabic source).

For the Brief Edits, Arabic/English editors from a professional translation house edited English MT

¹ We briefly experimented with having professional editors perform syntax-only corrections on the experiment corpus. This proved to be too difficult a task and did not show enough improvement to warrant its continuation.

output using the source Arabic as reference material. The editors used the company's existing guidelines for commercial post-editing work, which stressed accuracy over perfection, and called for translators to make only the changes required for understanding. The corpus was split in the same manner as for the Full Edits, with each editor working on documents from all three systems. Again, we analyzed 15 of the edited texts with the intentions of characterizing the types of corrections made and of feeding the results into an auto-correction tool. Unlike the Full Edits, all Brief Edit types were assessed as beneficial. For this reason, a new overall classification of edit types was found to be more appropriate. Of the 752 edits identified, 15% were placed in the category of easy to automate, 22% were categorized as difficult to automate, and 63% were categorized as impossible to automate or corrections should only be performed by a human. The majority category was changing/deleting one or more English words (e.g., arrestment to arrest) encompassing 26% of the total errors identified in the Brief Edits. This error type falls under the category of *should only be performed by a human*. We observed that the analysts were easily able to handle transliterated Arabic retained by the MT system and to "untranslate" Arabic names (e.g., *the lion* back to *Al-Assad*).

3.3 Machine Post-Edits

In our initial experiments, WordPerfect was the only COTS post-editor to have achieved a score significantly higher than that of the baseline MT output. For this reason, we performed a detailed analysis of the WordPerfect edits in both modes in which it was run. These included: 1) Grammar Only, in which the spell check feature was disabled and 2) All Features, in which the default settings, including the spell check feature, were enabled. Forty cropped documents were run in each mode and analyzed.

In the Grammar Only mode, a total of 126 edits were made to the set of 40 documents. Of these edits, 51% were in edit types classified as beneficial (e.g., changing verb form), 37% were neutral (e.g., merging or splitting words), and 13% were harmful (e.g., changing proper noun). The single largest edit type was changing a noun or adjective form (e.g., *businessmen* to *businessman*). This type of edit accounted for 30% of all edits and was catego-

rized as neutral. In the All Features mode, a total of 347 edits were made including spelling correction, which had an extremely deleterious effect on many proper names and all transliterated Arabic in the documents. Approximately 33% of edits were categorized as beneficial, 16% as neutral, and 52% as harmful. The largest edit type was changing an Arabic word or phrase. This type of edit accounted for 25% of all edits and was categorized as harmful.

Based on these findings, we designed an optimal WordPerfect configuration to feed into the analyst and translator experiments. This was a three-step process. In the first step, all features, or toggles, of WordPerfect's grammar checking component were identified and characterized. Out of 60 total toggles, 26 were found to be relevant for our purposes. These 26 features were then compared to the aforementioned edit analysis and speculatively classified as either beneficial or neutral. At this point, the MT output was processed using three different WordPerfect configurations: 1) only beneficial (Green, e.g., Infinitive) features were toggled on, 2) only neutral (Yellow, e.g., Hyphenation) features were toggled on, and 3) both beneficial and neutral (Green and Yellow) features were toggled on. In step two, the output from the first pass was analyzed and the features were redistributed according to the attested findings. This yielded two additional configurations: 1) new beneficial (Revised Green, e.g., Adverb moved to Yellow) and 2) new neutral (Revised Yellow, e.g., Verb Form moved to Green). The corpus was again processed using these two new configurations. In step three, the team examined the output of all five configurations and concluded that revised green settings created the optimal WordPerfect output and that this version would be included in the analyst and translator experiments.

4 Results

4.1 Analyst Results

All of the analyst analyses examined the means and standard deviations for acceptability ratings of the different tools. The specific descriptors chosen for the acceptability scale are quantitatively equidistant from each other based upon empirical data (Means, 2006), therefore the analysis of means, standard deviations, and t-tests are permissible. The acceptability ratings were coded into a scale of -3 (ex-

tremely unacceptable) to +3 (extremely acceptable). The means and standard deviations for all participant scores across the analyst questionnaire are illustrated in Figure 2.

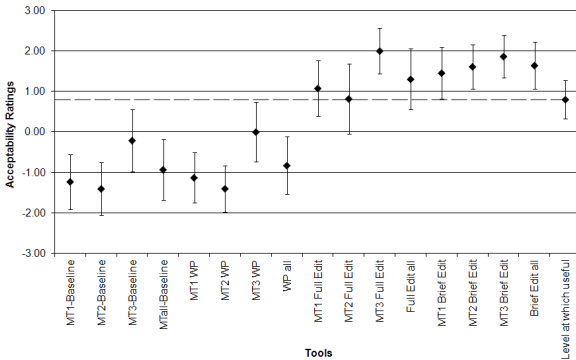
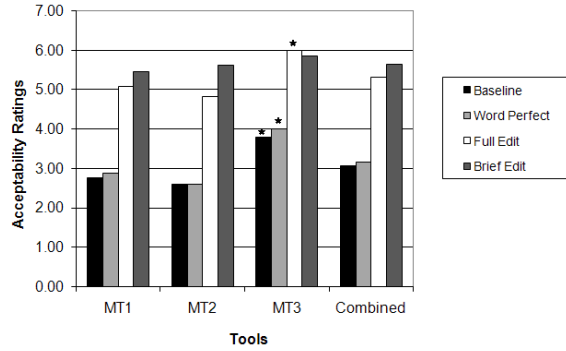


Figure 2. Mean Ratings.

In order to establish the level at which MT is useful to analysts, one of the final questions posed to participants in the Follow-on Questions section of the Analyst Experiment was “Thinking back to the rating scale, at what level of the scale does a document become useful to you?” The analysts answered this question based upon the same acceptability scale of “extremely unacceptable” to “extremely acceptable.” This level is shown as the dotted line in Figure 2. The figure shows that the human post-edited passages were rated at or above the Level at Which Useful, while the machine post-edited passages were rated lower.

4.1.1 Cross-MT Comparisons

Analyst ratings were compared across the three different MT systems in order to determine whether one of the engines produced passages that were more positively rated than the other two. Results from this comparison can be seen in Figure 3.



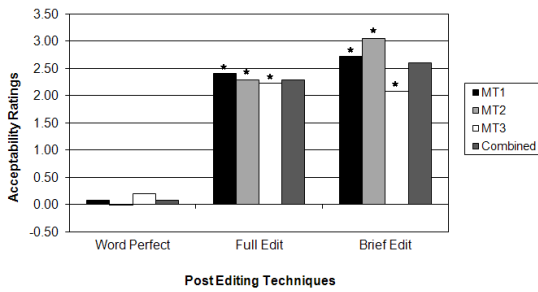
* Indicates a significant difference between MT3 and both MT1 and MT2.

Figure 3. Cross-MT Comparisons.

As seen in Figure 3, MT3 was rated significantly higher than both MT1 and MT2 in all categories but Brief Edit. While passages generated by MT3 were rated higher than the other MT tools, no significant differences were found between MT1 and MT2 ratings. To compensate for family-wise error (where the larger number of statistical tests increases the probability of incorrectly finding a significant difference due to chance) a modified Bonferroni procedure was used for this and subsequent analysis. This procedure reduces the normal significance level of .05 based upon the number of statistical tests.

4.1.2 Tools vs. Baseline

As illustrated in Figure 4, both the Full Edit and Brief Edit techniques produced passages that rated significantly higher than the raw MT (where the raw MT score is considered 0), but those produced by WordPerfect did not. This difference highlights the analysts’ preference for human post-edited passages as opposed to machine-based post-edited passages.



* Indicates a significant difference between the raw MT vs. Full Edits and raw MT vs. Brief Edits.

Figure 4. Tools vs. Baseline.

4.1.3 Tools vs. Level Useful

As mentioned earlier in this paper, at the end of the questionnaire, each analyst was asked to indicate the level at which they thought a passage would be useful. While this is a subjective rating, it provides a practical way to determine a baseline of acceptability against which the processed passage may be measured. As shown in Figure 5, the raw MT output and the machine-based post-edited output were rated significantly lower than the level useful; the ratings for Full Edits and Brief Edits were at least equal to the level determined useful; and all Brief Edits were rated above the level considered useful.

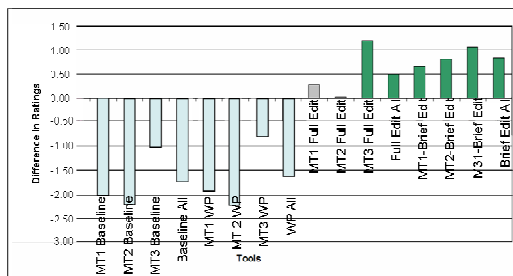
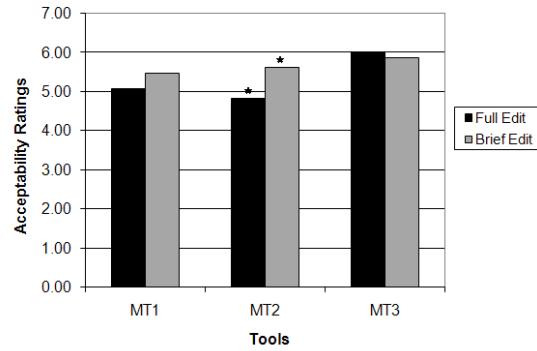


Figure 5. Tools vs. Baseline.

4.1.4 Full Edits vs. Brief Edits

While the results clearly show participants preferred the human-based post-edited passages, further analyses were performed to determine whether significant differences existed between the ratings for Full Edits as opposed to the ratings for Brief Edits.



* Indicates a significant difference between the ratings for Full Edit and Brief Edit on MT2.

Figure 6. Full Edits vs. Brief Edits.

As seen in Figure 6, there were no significant differences between ratings of the Full and Brief Edits performed on passages produced by MT1 and MT3. There was, however, a significant difference in ratings between the two human editing techniques with regard to MT2 output, with MT2 Brief Edits rating higher than MT2 Full Edits.

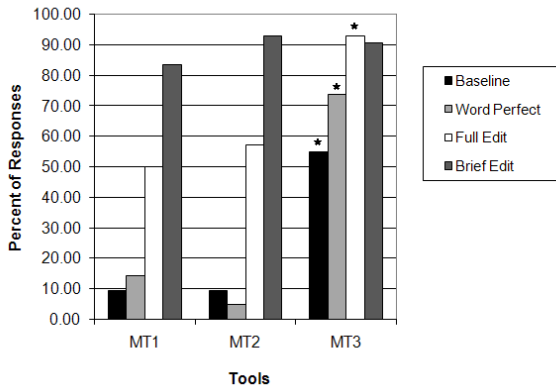
4.2 Translator Results

During the Translator Experiment, translators were asked whether they would “edit the translated document,” “use it as reference,” or “start a translation from scratch.” While the analyst acceptability scale used empirically derived equidistant descriptors (ratio data), the translator scales used non-numeric categories (categorical data). Therefore, the translator data analysis is somewhat different (i.e., Chi-square as opposed to t-tests) and does not allow for analyzing the differences between the means and standard deviations. The translator results are reported as either the percent of Edit responses or the percent of Edit plus Reference responses. The percent of Start from Scratch responses are not reported; therefore, the category totals do not always add up to 100%.

4.2.1 Cross-MT Comparisons

Translator ratings were compared across the different MT engines. Figure 7 illustrates that MT3 produced a significantly higher number of Edit and Reference responses than MT1 and MT2 in all but the Brief Edit categories. No differences were found between MT1 and MT2. These findings

concur with the differences found in the Analyst Experiment, which also rated MT3 passages higher.

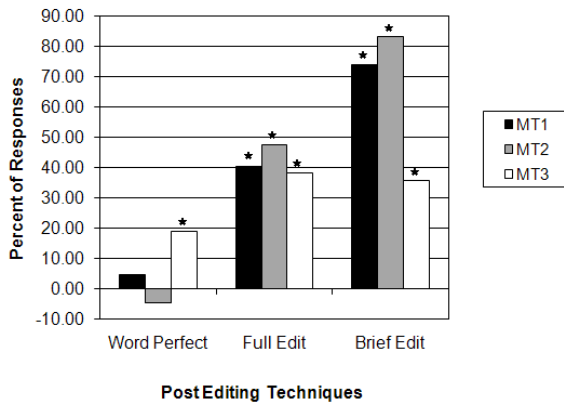


* Indicates a significant difference between the ratings for MT3 compared with MT1 and MT2.

Figure 7. Cross-MT Comparisons.

4.2.2 Tools vs. Baseline

Analyses were performed to determine whether the translators rated any post-editing techniques higher than the raw MT. The percentages of Edit plus Reference responses were used for this analysis.



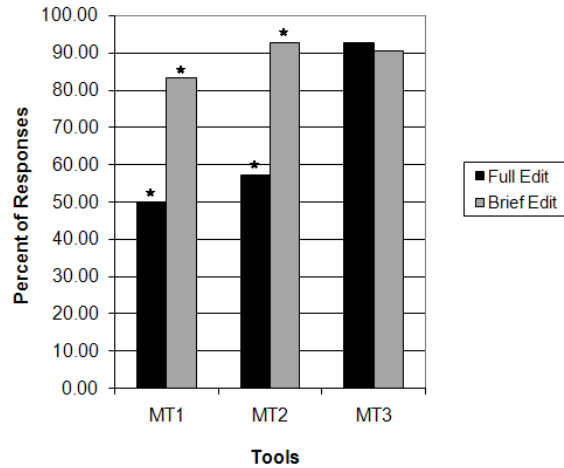
* Indicates a significant difference between the ratings when compared against the respective raw MT.

Figure 8. Tools vs. Baseline.

As seen in Figure 8, when comparing each post-editing technique against the raw MT score (where the raw MT score is considered 0), the human-based post-editing techniques (Full Edit and Brief Edit) were all rated significantly higher than the raw MT. In addition, MT3 post-edited by Word-Perfect received the highest rating among the MT engines.

4.2.3 Full Edits vs. Brief Edits

Similar to the analyst results, translators clearly preferred the human-based post-editing; therefore, further analyses were performed to determine whether significant differences existed between the ratings for Full Edits as opposed to the ratings for Brief Edits. Edit plus Reference responses were combined for these analyses.



* Indicates a significant difference between the ratings for Full Edit and Brief Edit for Edit + Reference responses.

Figure 9. Full Edits vs. Brief Edits.

Figure 9 illustrates that Full and Brief Edits of passages produced by MT1 and MT2 were rated significantly different from one another, with Brief Edits from both engines receiving more Edit and Reference responses. There were no significant differences between ratings for Full and Brief Edits of passages produced by MT3.

5 Conclusions

Our initial experiments with commercially-available automatic post-editors alone indicated that automatic post-editing performed by commercial tools neither improved the translators' or analysts' perception of the quality of MT output nor closed the gap between the ideal and current level of acceptability (usefulness level). A second round of assessment using a tuned commercial post-editing tool revealed that while automatic post-editing performed by machines remained below the translators' and analysts' perceived level of usefulness, all forms of post-editing performed by humans rated at

or above the perceived usability level. (It is worth noting that the tuned WordPerfect did show some measurable impact on acceptability.)

One way of interpreting these results is to conclude that analysts and translators are willing to use machine translation output once it has been post-edited by humans, using either the Full or Brief method of post-editing, and are not willing to use raw MT output or automatically post-edited MT output in its current state. However, these experiments were performed solely on output produced by Transfer-based, Arabic-to-English MT engines. It is possible that if these same experiments were run with different, more mature language pair output produced by MT engines of varying approaches, these conclusions could prove false. Nonetheless, it is fair to state that no acceptable automatic post-editor currently exists for Arabic-to-English MT.

Acknowledgments

The members of the Automated Machine Translation Improvement Through Post-Editing Techniques project would like to express our deepest thanks to our Government Sponsors and Arabic Subject Matter Expert (SME). Without their dedication, support, and academic expertise, the success of this project would not have been possible.

References

These are only a few of the most relevant references to this paper/project.

- Ariadna Font Llitjos and Jaime Carbonell. 2006. *Automating Post-Editing to Improve MT Systems*. Proceedings from the Association of Machine Translation in the Americas 2006 Conference (AMTA 2006) Workshop: Automated Post-Editing Techniques and Applications. Cambridge, MA.
- C. Donald Means. 2006. *Quantification of Response Alternatives of Acceptability, Adequacy, and Relative Goodness Ratings for a Working Age Population*. Unpublished Masters Thesis, University of Dayton, Dayton, Ohio.
- Greg Sanders. 2006. *Post-Editing in the GALE Program I*. Proceedings from the Association of Machine Translation in the Americas 2006 Conference (AMTA 2006) Workshop: Automated Post-Editing Techniques and Applications. Cambridge, MA.
- Jeffrey Allen. *What is MT Postediting?* A collection of reference material on MT Postediting. <http://www.geocities.com/mtpostediting/>
- John R. Pierce, John B. Carroll, et al. 1966. *Language and Machines — Computers in Translation and Linguistics*. Automated Language Processing Advisory Committee (ALPAC) report, National Academy of Sciences, National Research Council, Washington, D.C.
- Joseph Olive, Ph.D., DARPA GALE Program Manager. <http://www.darpa.gov/ipto/programs/gale/gale.asp>
- Julia Aymerich and Hermes Camelo. 2006. *Post-Editing of MT Output in a Production Setting*. Proceedings from the Association of Machine Translation in the Americas 2006 Conference (AMTA 2006) Workshop: Automated Post-Editing Techniques and Applications. Cambridge, MA.
- Marty Roberts. 2006. *Human Post-Editing of MT Output*. Proceedings from the Association of Machine Translation in the Americas 2006 Conference (AMTA 2006) Workshop: Automated Post-Editing Techniques and Applications. Cambridge, MA.
- Michelle Vanni. 2006. *Post-Editing in a Multi-Evidence Evaluation Paradigm: The PLATO Syntax Assessment*. Proceedings from the Association of Machine Translation in the Americas 2006 Conference (AMTA 2006) Workshop: Automated Post-Editing Techniques and Applications. Cambridge, MA.
- Rachael Allbritten. 2006. *Automated MT Improvement Through Post-Editing Techniques: Developing an MT Error Taxonomy*. Proceedings from the Association of Machine Translation in the Americas 2006 Conference (AMTA 2006) Workshop: Automated Post-Editing Techniques and Applications. Cambridge, MA.
- Stephanie Strassel. 2006. *Post-Editing in the GALE Program II*. Proceedings from the Association of Machine Translation in the Americas 2006 Conference (AMTA 2006) Workshop: Automated Post-Editing Techniques and Applications. Cambridge, MA.
- NIST: Open Machine Translation 2008 Evaluation (MT08). <http://www.nist.gov/speech/tests/mt/2008/>