Management of Large Annotation Projects Involving Multiple Human Judges: a case study of GALE Machine Translation Post-editing[*]

Meghan Lammie Glenn, Stephanie Strassel, Lauren Friedman, Haejoong Lee
LDC - Linguistic Data Consortium, Philadelphia, USA
{mlammie, strassel, lf, haejoong}@ldc.upenn.edu

Abstract

Managing large groups of human judges to perform any annotation task is a challenge. Linguistic Data Consortium coordinated the creation of manual machine translation post-editing results for the DARPA Global Autonomous Language Exploration Program. Machine translation is one of three core technology components for GALE, which includes an annual MT evaluation administered by National Institute of Standards and Technology. Among the training and test data LDC creates for the GALE program are gold standard translations for system evaluation. The GALE machine translation system evaluation metric is edit distance, measured by HTER (human translation edit rate), which calculates the minimum number of changes required for highly-trained human editors to correct MT output so that it has the same meaning as the reference translation. LDC has been responsible for overseeing the post-editing process for GALE. We describe some of the accomplishments and challenges of completing the post-editing effort, including developing a new web-based annotation workflow system, and recruiting and training human judges for the task. In addition, we suggest that the workflow system developed for post-editing could be ported efficiently to other annotation efforts.

1. Introduction

Machine translation is one of three core technology components for the DARPA Global Autonomous Language Exploration Program (GALE) Program, which includes an annual MT evaluation administered by National Institute of Standards and Technology (NIST). LDC creates training and test data for the GALE program, including gold standard translations for system evaluation. The GALE MT evaluation metric is edit distance, measured by HTER (human translation edit rate) (Snover, 2006). HTER calculates the minimum number of changes required for highly-trained human editors to correct MT output so that it has the same meaning as the reference translation (NIST, 2007). LDC has been responsible for overseeing the post-editing process for GALE Phases 1 and 2 (hereafter P1 and P2).

Manually annotating a large amount of data in a relatively short period of time (due to external constraints) poses a series of challenges extending to all aspects of the project. This paper focuses on work performed during GALE P2, describing LDC's approach to MT Post-Editing task definition, workflow, and editor recruitment. In addition, we will address some of the challenges LDC faced during this effort, such as coordinating multiple annotators working remotely, and balancing throughput with quality control and human management issues. As with every large scale annotation effort, the challenges posed by the project gave rise to interesting solutions and possibilities for future efforts, which this paper will also discuss.

2. Data Profile

The test data for GALE P2 included 60,000 words per language of Arabic and Chinese broadcast news and conversation, newswire, and web text. The three GALE research teams processed the test set, producing automatic speech recognition (ASR) where needed, and translating all data into English.

In addition to the 120,000 words of MT output contributed by each team, the post-editing test set included 12,000 words of translated data from GALE P1, so as to establish some comparability between the post-editing results for the two GALE phases. In all, there were 390,000 words of data to edit.

3. Project Overview

3.1 Task Description

GALE MT post-editing requires a human editor to compare a gold-standard translation to a system translation, modifying the system translation until its meaning is the same as the gold-standard reference. Editors work remotely, accessing post-editing assignments through a web-based workflow management site, which LDC developed for this task. Data is assigned in kits, or small folders which contain approximately 1200 words of translated material. Translations are reviewed by two independent editor "teams" of first- and second-pass editors. Additional quality control is performed by managers at LDC.

3.2 Task Definition

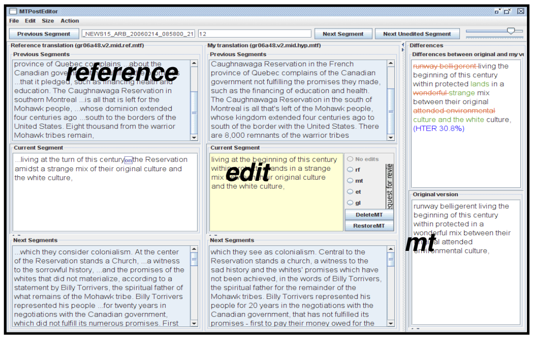
National Institute of Standards and Technology (NIST) and LDC collaborated to develop a set of post-editing rules, which describe in detail the goals of the task and instruct editors on how to handle specific aspects of the data. Using these guidelines, editors learn to make only necessary changes to the MT output, for example, by (1) adding meaning that is missing from the MT output; (2) removing extraneous material; and (3) shifting words and

[*] This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

phrases as appropriate, when the original placement obscures the meaning of the text.

3.3 Annotation Tool

NIST designed a customized java-based annotation tool, *MTPostEditor.jar*, shown in Figure 1, which displays the gold-standard reference in one column, the machine translation in another, and provides an editing area where editors make changes to the MT output. The tool is platform-independent and has been used with great success for post-editing in the GALE program.



_NEWS15_ARB_20060214_085800_21 | 12

Reference translation (gr06a48.v2.mid.ref.mtf)

Previous Segments

province of Quebec complains ...about the Canadian gove[illegible] ...that it pledged, such as financing health and education. The Caughnawaga Reservation in southern Montreal ...is all that is left for the Mohawk people, ...whose dominion extended four centuries ago ...south to the borders of the United States. Eight thousand from the warrior Mohawk tribes remain,

Current Segment

...living at the turn of this century on the Reservation amidst a strange mix of their original culture and the white culture,

Next Segments

...which they consider colonialism. At the center of the Reservation stands a Church, ...a witness to the sorrowful history, ...and the promises of the whites that did not materialize, according to a statement by Billy Torrivers, the spiritual father of what remains of the Mohawk tribe. Billy Torrivers represented his people ...for twenty years in negotiations with the Canadian government, which did not fulfill its numerous promises. First

My translation (gr06a48.v2.mid.hyp.mtf)

Previous Segments

Caughnawaga Reservation in the French province of Quebec complains of the Canadian government not fulfilling the promises they made, such as the financing of education and health. The Caughnawaga Reservation in the south of Montreal is all that's left of the Mohawk people, whose kingdom extended four centuries ago to south of the border with the United States. There are 8,000 remnants of the warrior tribes

Current Segment

living at the beginning of this century within protected lands in a strange mix between their original culture and the white culture,

Next Segments

which they see as colonialism. Central to the Reservation stands a church, a witness to the sad history and the whites' promises which have not been achieved, in the words of Billy Torrivers, the spiritual father for the remainder of the Mohawk tribes. Billy Torrivers represented his people for 20 years in the negotiations with the Canadian government, that has not fulfilled its promises - first to pay their money owed for the

Differences

Differences between original and my ve

~~runway belligerent~~ living the beginning of this century within protected lands in a ~~wonderful~~ strange mix between their original ~~attended environmental~~ culture and the white culture, (HTER 30.8%)

Original version

runway belligerent living the beginning of this century within protected in a wonderful mix between their original attended environmental culture,

Figure 1: Machine Translation Post-Editing tool

Documents are edited one sentence at a time. Editors are able to view the differences between their edits and the MT output, and to see the HTER score for each sentence. They are also able to read the entire document in this tool, so as to better understand the working sentence in context.

3.4 Editor recruitment and training

LDC recruited Philadelphia-area native English speakers with training in copy-editing, proofreading, creative writing, or journalism. All candidates received a pre-test kit, or file archive containing a brief set of instructions, the annotation tool and the sample file to edit. The test file consisted of 10 sentences, selected from GALE P1 data, which demonstrated a range of genres, source languages, and machine translation systems. Sentences were also selected for a range of editing difficulty.

Managers scored and reviewed the test kits carefully, looking for a basic post-editing aptitude in the test kit responses: applicants who did not add extraneous information to their edits, who spelled words correctly, and who incorporated the full meaning of the gold-standard in their edits. After examining the pre-test results and eliminating outliers, we invited qualified applicants to LDC's office in Philadelphia for an intensive training session.

The training session focused on the MT post-editing guidelines, and displayed a set of examples of possible edits. Following the training session, applicants re-edited the test kits so that managers could observe what they would do differently after learning more about the task. Those who continued to produce edits that conveyed the same meaning as the gold-standard translation and who made only necessary changes to the MT were selected for the project.

Before beginning work on GALE P2 production data, editors read the guidelines carefully and completed a starter kit. The starter kit reinforced their knowledge of the post-editing rules and allowed editors and LDC staff to solve technical problems. It also offered managers another opportunity to evaluate the editors, and to answer many task-related and procedural questions before the project started in earnest.

3.5 Kit composition

LDC worked closely with NIST to draft a plan for kit size and assignment order. At minimum, a kit is a folder containing the human reference translation and the machine translation that the editor will compare and edit. It might also contain guidelines, the annotation tool, or other documentation, such as a list of examples. GALE P2 kits contained gold-standard reference and MT output files of approximately 1200 words each, or about 3-4 documents of 200-250 words. To ensure objective evaluation of each team's submissions, NIST devised a Latin square to mix the data from each research team, source language, document length, and genre, organized the documents into kits, and established the kit assignment order based on the Latin square design. In general, a kit of 1200 words takes about 3 hours for an editor to evaluate.

3.6 Workflow design

In order to improve editor accuracy and reduce the impact of outlier edits, the project design required multiple independent reviews of the MT data. Every version of the data was edited by four editors: two first-passes, which were then checked by a second pass. Editors were assigned the role of first or second pass at the beginning of the project, and in general, retained that role for the duration of the project. Second pass editors were assigned to first pass editors, to form a team. In GALE P2, editor teams were assigned 36 kits each. Figure 2 illustrates how a kit progressed through LDC's workflow system:

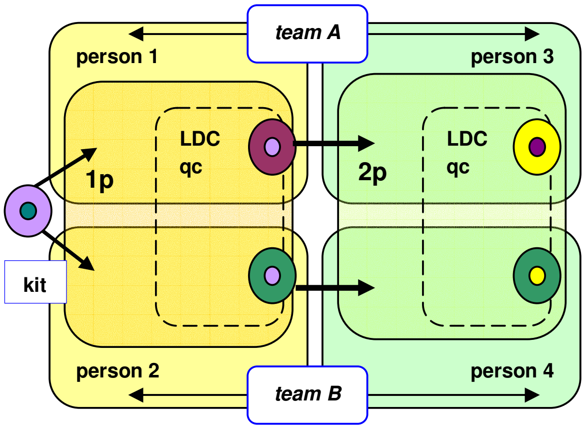


Figure 2: LDC MT Post-Editing Workflow.

3.7 Workflow management system

Files were managed by a workflow system, designed specifically at LDC for this task. Assignments for each editor "team," or first- and second-pass pair, were loaded into the system at the start of the project. LDC managers were able to view users, manage assignments, and backup the project through this system, as appropriate.

Editors checked out kits from this system, working with one kit at a time. After a first pass editor checked in a completed kit and there were no problems identified with it by an automated scoring process, the kit is automatically assigned to the second pass editor.

The workflow system was designed as a central information resource for editors, as well. Here they can view their file assignment lists, summaries of their expected payment per kit, and links to other project resources. Editors are also able to check the status of their first- or second-pass partner, in order to manage their time more efficiently.

Table 1 shows a summary of the number of editors and the data reviewed during this effort

GALE P2 MT Post-Editing	
Number of editors	36
Number of kits (after 4 reviews)	1300
Number of words (after4 reviews)	1,560,000
Total number of editing decisions	617,000

Table 1: Post Editing Data Volume Summary

3.8 Quality control

The first stage of quality control is the second pass. The first pass editor's goal is to make as few edits as possible to match the MT output meaning to that of the gold-standard reference. The second pass editor has an added layer of responsibility: to check the meaning again, to reduce edits where possible, and to correct careless errors.

In addition, a number of automatic and manual quality control mechanisms are in place at LDC during the editing process to catch careless errors or alert managers to potential problems. For example, the first kit submission of every editor is marked broken and is held in a separate queue until approved by a manager. Scripts automatically score and check incoming kits to flag potentially problematic kits. These include kits with high scores or with unedited segments. Managers also spot-check kits daily, and provide feedback to editors accordingly.

4. Project management

Maintaining a consistent level of understanding and practice with a large group of annotators requires frequent contact with each person, and frequent review of key principles. After the initial face-to-face interviews and training sessions, editors worked remotely and were not privy to on-site meetings to resolve frequently asked questions or correct misunderstandings. Therefore, LDC's post-editing management team provided constant feedback to editors over email and through the online workflow management system in order to satisfy the quality requirements for the project.

4.1 Task challenges

While the primary rules of post-editing are relatively straightforward, the practice of post-editing can be very difficult and very tiring. Editors strive to retain as much of the original MT output as possible, adding, moving, or inserting information only if the meaning of the MT does not match the meaning of the reference. For example:

Gold-standard translation: OK, very nice.

MT output: The in

With this system translation, the editor must delete the MT output and insert the shortest possible phrase, such as, *OK nice* or simply, *Nice*. Sometimes the MT output is simply not English and the editor will take the same approach as in the previous example: replace the original output with the shortest possible phrase that means the same as the gold-standard reference.

In addition to linguistic issues, editors may encounter challenging display issues. As shown in our description of the annotation tool, gold-standard reference translations and MT output are aligned at the sentence level and are loaded into the post-editing tool simultaneously. Editors step through a document one sentence or phrase at a time. However, because alignment of system and reference translations is automatic, errors may occur. For example,

Gold-standard translation: Is there any link between achieving democracy and social life?

MT output, segment 1: And, democracy and social life What is

MT output, segment 2: the connection,

In such instances, editors do not move words from the second MT segment up to the first; instead, they must think creatively beyond physical divisions in the texts and in the tool, and must train their focus to the whole document. In this case, the editor would leave "the connection," in segment 2, such as in this edited example:

Human edit, segment 1: And, between achieving democracy and social life, is there

Human edit, segment 2: any connection?

4.2 Other project management tools

The website and workflow LDC designed for this period included a number of support mechanisms to give editors immediate feedback on certain quality issues and to provide a forum to ask questions, monitor progress, and even blow off a little steam. These support mechanisms are described in the sections that follow.

4.2.1 Request Tracker

LDC managers and support staff used the Request Tracker (RT) system[1] to communicate with large groups of editors. We also encouraged editors to contact us through the RT system so that multiple managers would be able to respond in a timely fashion to the editor's question or concern. The RT system also received notifications when an editor created an account, when an editor submitted his or her first kit, and when a kit failed to complete the scoring process.

4.2.2 Comment function

A convenient feature of the web-based workflow management system is a log for editors to post anonymous comments about their assigned kits. The log also documented any change to a kit (when an editor checked out the kit, etc.) with a preformatted comment. The comments remain on a web-page for the kit, so that

[1] http://bestpractical.com/rt

when a manager leaves a comment, both the first and the second pass editors can review it.

The editors who used the comment system to request feedback and to chat with their partners seemed to benefit from the exchanges. Careful review of the comments from editors is both interesting and informative. In a few cases, we noticed a friendly bond develop between editors. In other cases, editors warned one another about difficult regions of the data, or discussed editing rules. Many of the comments are related to scheduling issues; comments second pass editors were often to ask for more data from their first pass partners.

5. Conclusion

In this paper, we have described the process that LDC implemented in order to successfully complete GALE P2 MT Post Editing effort, including data volumes, task definition, workflow design, and management tools. The approach described here worked particularly well for a large annotation project involving multiple human annotators. We were encouraged by the success of the system, and are currently planning extensions to the infrastructure, such as facilitating a more efficient ramping up of the system for new annotation efforts. The tools designed for this effort are by no means restricted to a post-editing project, but could be used for any project involving the outsourcing of data to remotely-located individuals.

6. Acknowledgements

This work has been supported under the GALE program of the Defense Advanced Research Projects Agency. We would like to thank Mark Przybocki, Audrey Le, and Sebastien Bronsart at NIST for annotation tool development, kit creation, for scoring post-editing results, and for collaborating with LDC on the annotation guidelines and workflow design. We also thank the LDC support staff, without whose help the GALE P2 post-editing effort would not have been completed on time. We are also thankful to Shawn Medero, who built the GALE P2 workflow management system while at LDC.

7. References

NIST, LDC (2007). Post Editing Guidelines Version 3.0.2 http://projects.ldc.upenn.edu/gale/Translation/Editors/GALEpostedit_guidelines-3.0.2.pdf

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. (2006). "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas.