[From: *Translating and the Computer 30*, November 2008 (London: Aslib)]

A Quality Control Framework for Gold Standard Reference Translations: The Process and Toolkit Developed for GALE[1]

Lauren Friedman, Haejoong Lee, Stephanie Strassel
{lf, haejoong, strassel}@ldc.upenn.edu
Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA, USA

Abstract

While a multitude of machine translation (MT) evaluation metrics exist, most require one or more gold standard references. For the DARPA GALE program, source data is translated according to detailed guidelines[2] and high quality standards, but these raw translations then undergo a rigorous and carefully constructed quality control (QC) process to create the final references. GALE evaluation translation references undergo several distinct phases of translation and quality control, utilizing a specially-designed toolkit, QCTool. This paper describes this framework and suggests that it may be successfully applied to improve and streamline other large-scale translation projects.

1. Introduction

While a multitude of machine translation (MT) evaluation metrics exist, most require one or more gold standard references. For the DARPA GALE program, source data in four genres (newswire, web, broadcast conversation, and broadcast news) is translated according to detailed guidelines[2] and high quality standards, but these raw translations then undergo a rigorous and carefully constructed quality control (QC) process to create the final references.

Creating human translations (HT) for the express purpose of MT evaluation requires different standards and priorities than HT created for general use. GALE evaluation translation references undergo multiple distinct phases of translation and quality control. Raw translations are created by translators under contract to the Linguistic Data Consortium

[1] 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.

[2] GALE Translation Guidelines, Version 2.4. http://proiects.ldc.upenn.edu/gale/Translation/

(LDC). Translators follow LDC's GALE translation guidelines that include rules for handling idioms, non-standard grammar, misspellings, ambiguity, proper names and other common complexities, as well as genre-specific issues like URLs and emoticons in web text.

While traditional translation tasks and some machine translation evaluation protocols would accept the output of the raw translation stage as adequate, GALE requires reference translations to undergo several additional stages of annotation and quality control to correct errors, improve translation adequacy, add translation variants, standardize proper nouns, verify technical terms and so on, with the ultimate goal of having the gold standard translations that are absolutely faithful to the source in terms of meaning, fluency, structure and style.

The resource-intensive process for gold standard creation, which utilizes a specially-designed toolkit, QCTool, was developed in part to meet GALE's need for multiple editing passes on translation data wherein each intermediary version of the translation is preserved. This paper describes this framework and suggests that it may be successfully applied to improve and streamline other large-scale translation projects.

2. The QC Process

The quality control process for GALE evaluation references was designed to mediate the variation in quality and consistency that always occurs in human translation. Even when translations are created by skilled translators and then proofread closely, erroneous translations and areas of ambiguity frequently remain. Certain genres that are evaluated in GALE, especially broadcast conversation and web data, provide particular challenges to human translators as well as to the MT systems being evaluated. In addition, since translation is not an exact science, two independent translators will typically create two different translations for an identical source document. However, for evaluation reference data, the translation must be exact and fully expressive: accuracy is emphasized, even when it comes at the expense of fluency, and the translation must convey no more and no less than the source document.

In order to standardize the translations and produce an appropriate reference for evaluation, LDC has developed a six-step translation and QC process:

1) source-language dominant bilingual translator produces a preliminary translation emphasizing accuracy;
2) target-language dominant bilingual translator revises the translation to improve fluency;
3) source-language dominant bilingual annotator checks translation for errors and omissions;
4) source-language dominant bilingual senior annotator checks for remaining errors, Improves fluency, corrects and standardizes named entities;
5) target-language dominant bilingual annotator improves fluency and adds translation variants where required;
6) target-language monolingual annotator reviews for fluency and consistency, and flags questionable regions.

Steps 1 and 2 are outsourced to commercial translation agencies that have been vetted by LDC. These agencies are required to follow LDC's translation guidelines[2] to the letter, with additional quality control loops added to meet the evaluation data standards. The translations that they deliver after Step 2 are considered final and complete by the agencies; the subsequent steps are an above-and-beyond layer added by the LDC in order to ensure the highest possible confidence in the released gold standard.

Steps 3 through 6 are performed in house. In Step 3, the annotator focuses only on correcting egregious errors. The main objective of Step 4 is to smooth out any more nuanced issues with the translation, while verifying total fidelity to the source - a requirement for the GALE evaluation. In Step 5, fluency problems are ironed out and translation variants are introduced to clarify regions of ambiguity.

Relative to Step 5, Steps 3 and 4 are extremely time-intensive. For the most recent GALE evaluation, Step 3 averaged 25 minutes per Arabic document and 21 minutes per Chinese document. (Each document is between 150 and 250 tokens.) Step 4 averaged 13 minutes per Arabic document and 15 minutes[3] per Chinese document. Step 5, however, averaged just 6

[3] These figures are from two evaluations ago since - in the most recent evaluation - there were some issues in recording time-per-file in Steps 4 and 5.

minutes[3] per file for both languages, since the translation is generally in excellent shape by the time it reaches Step 5.

The final check, Step 6, is a quick but thorough read-through of all of the translations, with an eye to any errors that may have been inadvertently introduced in previous steps. The Step 6 reviewer must also ensure that the translations read as correct, fluent English, independent of the source text.

Each stage requires a different level of expertise, and while six different translators might produce six different translations, the six-person translation and QC team - who work independently but consecutively on one working document - is designed to achieve a higher level of consistency and predictability. Step 5, where translation variants are introduced when required, is especially important for ensuring that any alternate - but equally accurate - interpretations of the source text are included in the final translation reference.

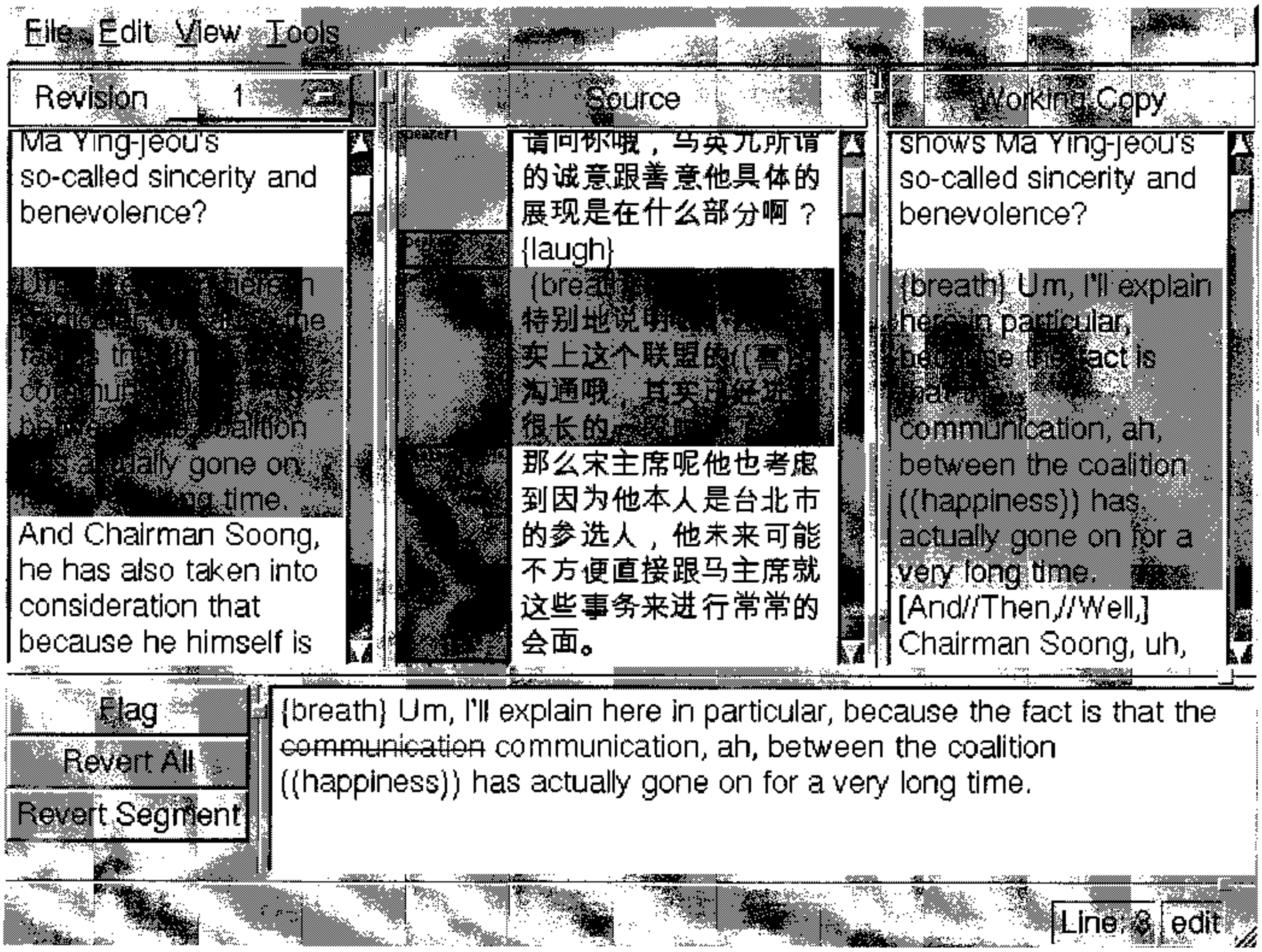
3. QCTool

The multi-stage QC process is facilitated by QCTool, an annotation tool written in Python and based on XTrans[4], LDC's specialized transcription tool. QCTool allows annotators to view source documents and translations side-by-side and edit their working copy of a translation. It also includes functionality for viewing and reverting to previous translation versions, flagging sections for further review, playing back original audio data (where applicable), and displaying edits as they are made (see Fig. 1).

QCTool was developed by reusing and specializing components of XTrans. The amount of newly added code was minimal. A simple analysis shows that QCTool reused 99% of the XTrans code. About 8% of the QCTool code was newly added, mostly by specializing components of XTrans through subclassing. For example, the text display had to be enhanced to make sure that corresponding segments are aligned when displayed on the tree-panel text display. This approach made the rapid development of the tool possible.

[4] Glenn, M. L. and Strassel, S. “Shared linguistic resources for the meeting domain.” 2006. *Proceedings of the 2006 CLEAR and RT Evaluations.*

QCTool uses a slightly modified version of the original data model used by XTrans. The original data model is basically a table where each row represents a segment or sentence. The translation QC process involves several layers of such tables. For example, source text, preliminary translation and corrections from further QC steps each form a layer. In QCTool, these layers of tables are combined into one table by means of table ID, i.e. each row is augmented by an ID of the table it belongs to. The new data model is physically stored using the same file format used by XTrans. However, only QCTool recognizes the additional information to display each row on an appropriate text panel.



Ma Ying-jeou's so-called sincerity and benevolence?

[illegible] coalition has actually gone on, [illegible] time.

And Chairman Soong, he has also taken into consideration that because he himself is

请问你哦，马英九所谓的诚意跟善意他具体的展现是在什么部分啊？{laugh}

{breath} [illegible] 特别地说明 [illegible] 实上这个联盟的((高)) 沟通哦，其实已经进 [illegible] 很长的 [illegible]

那么宋主席呢他也考虑到因为他本人是台北市的参选人，他未来可能不方便直接跟马主席就这些事务来进行常常的会面。

shows Ma Ying-jeou's so-called sincerity and benevolence?

{breath} Um, I'll explain here in particular, because the fact is that the communication, ah, between the coalition ((happiness)) has actually gone on for a very long time.

[And//Then,//Well,] Chairman Soong, uh,

{breath} Um, I'll explain here in particular, because the fact is that the ~~communication~~ communication, ah, between the coalition ((happiness)) has actually gone on for a very long time.

Figure 1. QCTool showing a portion of a Chinese broadcast conversation document. The left column shows the previous version of the translation (which can be toggled using the "Revision" dropdown); the center column shows the source text; and the right column shows the editable working copy. The bottom pane shows edits as they are made.

As with LDC's other annotation tools, QCTool is fully integrated with the Annotation Workflow

System (AWS)[5], which manages documents, directories, permissions, and assignments, as well as tracking annotator efficiency and progress. Together, QCTool and AWS ensure that each intermediary version of the translation is stored for later training and analysis. In LDC's production pipeline, QCTool is integrated with AWS for steps 3, 4 and 5 of the six-step process. Before entering the AWS workflow, the file contains two layers of text: source text and a preliminary translation. At each step of the AWS workflow, AWS adds a new layer, which is an exact copy of the previous translation layer. The annotator at each step works on the newly added layer, which is displayed on the "Working Copy" panel, to correct and improve it. One of the previous translations is displayed on the "Revision" panel, and user can select which revision to display. The source text is displayed on the "Source" panel.

QCTool increases the speed and accuracy of the quality control process by giving the annotator access to all of the relevant information in one place, and making it much easier to visualize revisions and zero in on problematic regions.

4. Conclusion

For MT evaluations, where an MT system is penalized when it differs from the reference, a process that endeavors to produce a fully correct and comprehensive translation is essential, and it was with this goal in mind that this framework was developed. The observed advantages of the framework described in this paper include improved translation accuracy, consistency, and fluency, as well as increased annotator efficiency. However, it is important to note that the costs - in time, training, and funds - are great, and this resource-intensive procedure only makes sense when even small errors would be extremely problematic in the final product.

While the QC process and tool were developed to support the GALE MT evaluation, this QC framework can potentially streamline and standardize the translation process for any project where quality and efficiency are paramount, or where intermediary translation versions must be archived.

[5] Glenn, M. L. and Strassel, S, 2006.