Beyond MT: Source Content Quality and Process Automation

Jenny Lu
CA Technologies
Islandia, NY 11749
Jenny.Lu@ca.com

Patricia Paladini Adell
CA Technologies
Barcelona, Spain
Patricia.PaladiniAdell@ca.com

Abstract

This document introduces the strategy implemented at CA Technologies to exploit Machine Translation (MT) at the corporate-wide level. We will introduce the different approaches followed to further improve the quality of the output of the machine translation engine once the engines have reached a maximum level of customization. Senior team support, clear communication between the parties involved and improvement measurement are the key components for the success of the initiative.

1 Introduction

As the deployment of machine translation matures, the never ending requirement of reducing translation cost and increasing productivity remains the main goal in the organization.

At CA Technologies, the MT engines are owned and maintained by the internal translation team. Currently we work with a mixture of rule-based (RBMT) and statistical (SMT) machine translation engines depending on the languages:

- Lucy LT: this RBMT is used to machine translate FIGS
- Moses: this SMT is used for Portuguese Brazilian and Italian
- The Toshiba The Honyaku Server Enterprise Edition (The 翻訳®)* (RBMT) is used to MT into Japanese
- The CCID Intelligent Translation System (赛迪智能翻译系统) (RBMT) is used for Simplified Chinese

*The 翻訳® is a registered trademark of Toshiba Solutions Corporation.

For Nordic languages, we have some ongoing projects to train Moses and we are also doing some preliminary work on Hybrid engines for different languages. Our current MT process is applied to the localization of both, user interface (UI) elements (software strings) and product documentation.

MT was implemented back in 2009. Since then, CA translators have fine-tuned and enhanced the MT constantly. Today, we believe that our engines are fully customized to our needs and only minor routine maintenance tasks are performed. However, we believe that the machine translation output has not reached its maximum quality level, regardless of how well the MT engines perform. Over the years, it has been reported by the translators that the source content is not MT friendly and the quality of the MT output could be enhanced by providing MT-friendly source content. Backed up by the data collected in the translation query system and with the support of the senior management team, the localization team reached out to the technical writing team as well as the development team to address the issues with the goal to help those teams to provide MT-friendly content for localization.

The immediate approach was to examine the source input to the MT engines and the source file turnover process for localization for both, product user interface files and documentation. After that, several improvement processes were put in place in the area of product translation: Create an end-to-

end solution to process software translation including automation of machine translation to the software files, apply controlled English authoring to the product documentation, and define clear measurement of the productivity improvement. Details will be provided in the following sections of this paper.

All these initiatives were included under the umbrella of a more ambitious strategy that consisted shifting the ownership of releasing localization products to development teams. This strategy, called "Island to Mainland", proposed the integration of the localization process into product development main stream. The initial effort required different tactics when dealing with tech writing and development. However, the concept of moving the localization ownership from "Island" to "Mainland" works the same. Senior management teams' support was critical for the success of the proposal.

2 Product UI translation

In the field of UI translation, to support "Island" to "Mainland" goal and to help development team to provide localization friendly resource files, an end-to-end solution was implemented to handle resource files turned over for translation. Today, the fully automated localization process is kicked-off by the development team uploading source files into a home grown workflow system designed to manage the end-to-end UI translation. The workflow includes several automated steps that guarantee a seamless localization process: starting from i18n testing and source file acceptance, pre-process of the source files, automatic TM leverage, automatic MT application, post-editing by the translators, post-process of the translated files, to the final check-in of the translated source files.

Regarding the i18n testing and acceptance phase, the system performs a source texting for i18n and L10n issues and provides a report with the issues identified and several remediation proposals to be implemented by developers. The system also includes a pseudo-localization utility and an English spell and grammar checker that help to identify non-MT friendly source text. We continue work on the proposal to reject those UI packages turned over for localization that do not meet a minimum L10n and i18n quality score.

Once the files are i18n and L10n issue-free, the files are automatically parsed by the UI translation workflow system (also a home-grown tool) and the text is leveraged against a translation memory repository. After the translation memories have been applied, any output string that is neither a full match (100% match) or a fuzzy match, will be automatically processed through Machine Translation. The resulting file (a mixture of 100% matches, fuzzies above 70% and machine translated strings) will be ready for post editing by the internal translation team.

At this point, an automatic notification will be sent to translators with a translation request. All they have to do is to log into a home-grown translation editor and check out the files for post-editing.

Once the post-editing is completed, files are automatically checked back into the workflow system and post-processed by the system, and the localized files are uploaded into development repository automatically. The translation memories are updated at the same time.

3 Product documentation translation

Several initiatives have been implemented in view of enhancing the quality of the MT output, most of them aimed to enhance the quality of the English source text. The localization team first identified the communication challenges among tech writers and localization, set priorities for the different challenges, and agreed on corrective actions.

To improve communication among both teams, the turn-over form used to turnover files for localization was enhanced with additional information useful for translators, like product training information, related products information, location of third party texts which are not to be translated, etc.

A significant effort was also done on terminology management by moving English terminology management ownership to technical writers in order to release translators from the task of extracting terminology out of the files turned over for translation. To achieve this goal, both teams have defined new process to manage terminology that implies technical writers to feed an English online dashboard as they write with new terms and definitions. Once the terms are added into the dashboard, localization team provides translations to the different terms and RBMT engines dictionaries are

updated, if applicable. As a result, technical writers acquired new terminology management tools compatible with their authoring system.

On the controlled authoring side, localization team helped tech writing team to identify grammar structures that were not MT-friendly and, thus, were provoking a poor MT output that needed a significant post-editing effort. The controlled authoring tool was enhanced with these rules and technical writers were instructed to accept the remediation offered by the tool. One more strategy to support controlled authoring was to design a process to automatically report how MT-friendly the source file is based on the MT-friendly rules added to the controlled authoring tool. The resulting index (poor, acceptable, excellence) is currently included in the turnover form. Same as for the UI files, we are working on a proposal to reject those files turned-over for localization that do not meet a minimum MT-friendly content score.

On top of all the strategies described above, several training was provided to tech writers on how to write MT-friendly content and how to use source control authoring tools.

In addition, tech writers have designed a new strategy to write source content that is proving to be very useful to improve the quality of the MT output. This strategy consists of moving away from the traditional A to Z user guide that includes all the text in long chapters and paragraphs and focusing on short descriptions of the different task end users have to perform. The result is a source text containing short instructive sentences that have been proven to be a MT-friendly input.

In order to support the implementation of all the proposals, the query and issue tracking system used by CA translators to report source issues found during post-editing was also enhanced to allow translators to easily report source content issues by categories. The items reported are categorized between queries (for example, need technical clarification) and issues that are affecting the quality of the MT output. The issues would include typos, English inaccuracies, grammar issues, translatability issues (whether strings should be translated or not), incorrect use of abbreviation and acronyms, etc.

As a follow-up on progress and to provide updated measurement, we provide issue reports on a project base to help tech writers to identify MT-friendly issues and to take the necessary corrective actions.

4 Measurement

Keeping measurement data has been a general practice within the translation team at CA Technologies over the past few years. Besides tracking the number of queries and issues raised by translators during translation and the categories of those queries, which helps the team to provide facts to the tech writing team on the quality improvement of English content, the translation team also developed its productivity tracking system which tracks translation and MT volume, post-editing throughput, and rework time. The data collected is then used to create trends and helps tech writers and developers identify areas of improvement needed.

Example of measurement includes benchmark data taken at the beginning of the improvement initiatives:

Category of queries	Total #	%
English Inaccuracies	796	21.97%
Translatability	561	15.48%
English Language Correctness	555	15.32%
Context	534	14.74%
Abbreviations/Acronyms	302	8.34%
Technical Clarification	296	8.17%
Placeholders/Tags	218	6.02%
Others	361	9.96%
Grand Total	**3623**	**100%**

Table 1: Top categories of issues in the source files reported by translators during SW localization

Category of queries	Total #	%
Translatability	493	22.59%
English Language Correctness	417	19.11%
English Inaccuracies	389	17.83%
Cross-Reference Unfound	354	16.22%
Cross-Reference Inconsistencies	165	7.56%
Others	364	16.69%
Grand Total	**2182**	**100%**

Table 2: Top categories of issues in the source files reported by translators during doc localization

Example of measurement by products:

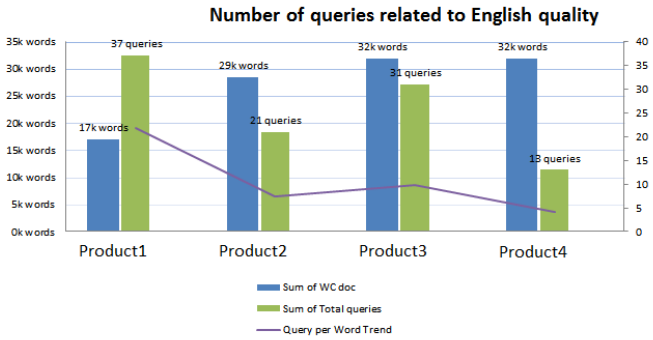

Table 3: Issues reported by products

5 Additional initiatives

Ongoing MT initiatives at CA Technologies include: development of Nordic language engine; improve current MT capacities by designing MT hybrid approach for certain languages, e.g. Japanese to Korean; designing hybrid approach (RB and SMT engines) for several languages; providing real-time MT translation for non-static and user created content such as forums, web sites, user communities, etc.; and expand MT capabilities by using both SMT and RBMT.

Other initiatives not directly related to MT include a process to eliminate ambiguity of UI strings due to missing context in the UI strings to reduce UI validation cycles and a process to tag UI elements in documentation to educate MT engines to ignore UI elements while providing a MT proposal.

6 Conclusion

In order to further improve the quality of MT engines once they have reached a maximum level of customization, the quality of the source content must be MT-friendly and the source file format must be standardized to avoid manual process prior to MT process. In order to fix the fundamental issues, and to secure the quality of the source content, a corporate-wide strategy is necessary to foster the communication between localization team and the source owners, technical writing and the development community. The initiatives were well received by the community and new processes and tools were implemented to improve the source content creation.