Machine Translation (MT) @ CA Technologies: Where can MT be best successful and what are the best MT engines for various languages

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

Abstract

CA's globalization team has a long term goal of reaching fully loaded costs of 10 cents per word. Fully loaded costs include the costs incurred for translation, localization QA, engineering, project management, and overall management. While translation budgets are gradually decreasing and volumes increasing, machine translation becomes an alternative source to produce more with less. This paper describes how CA Technologies tries to accomplish this long term goal with the deployment of MT systems to increase productivity with less cost, in a relatively short time.

1 Introduction

CA Technologies started its first series of trials with four MT engines in 2004-2005, including both RBMT and SBMT. The trials were performed at the headquarters with little involvement from the linguists except for quality reviews. The languages tested included French, German, Spanish, and Japanese. The trials were disrupted by the unstable quality and intensive labor required to carry on the trials. In 2008, CA took a different approach to empower the in country linguist teams to take over the research, investigation and vendor negotiations. Different language MT systems were purchased and implemented within six months, except Japanese which took one year. Currently, the linguist teams keep the ownership of maintaining and supporting the systems locally. The original idea was to apply MT to the documentation translations due to the high volumes. However, through the continued exercise, the linguist teams determined that MT is also suited for UI translations due to their short sentences. Today, MT engines are in full production at CA for UI and documentation translations in Japanese, S. Chinese, B. Portuguese, French, Italian, German, and Spanish.

The Japanese team took the quickest opportunity of implementing a Japanese RBMT engine with support from CA's globalization engineers and architects. The S. Chinese team and the FIGS teams were able to leverage the experiences from the Japanese system setup and followed up with their own RBMT system. B. Portuguese was behind at first, but caught up and implemented a SBMT as part of a study in the globalization team in India. In the case of B. Portuguese, the two teams – linguist and engineering – collaborated on testing the engine using legacy TMs, and were able to produce good-for-post-edit quality output.

Today, MT translation plus post-editing of product and documentation are fully in production for the six languages. The translation throughput can be increased up to 100% in some languages, and the outsourcing cost of post-editing is reduced by 30%-50% with the potential for additional savings.

2 MT engine selection

The MT engine selection process for each language was focused on:

- Translation Quality
- Dictionary Usability
- Features Availability

Tests were performed with various types of documents with and without dictionaries. The quality evaluation was done with a simplified method. It is worth to mention that system support from the internal engineering team was vital to the success of the system implementation. Due to the nature of the translation tools used by the linguist team, integration between the MT engines and the translation system required customized support depending on each MT engine. CA Globalization's architect team provided pre and post processing tools to solve the issues unique to each MT engine by working with the linguists directly.

The following MT engines are implemented at CA:

- Japanese - Toshiba The Honyaku (The 翻訳®)*
- S. Chinese - CCID Intelligent Translation System (赛迪智能翻译系统)
- French - Lucy LT
- German – Lucy LT
- Spanish – Lucy LT
- Italian – Lucy LT and Moses
- B. Portuguese – Moses

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

3 Strategy of MT implementation at CA

These are the strategies applied during MT implementation:

- MT systems implemented, maintained, and owned within local countries
- Mixture of rule and statistics based engines depending on the language
- Dictionaries maintained by local linguist teams
- Engine improvement monitored by local linguist teams
- Post-editing by internal teams and outsource
- Engineering team supports system integration

4 Factors to success

The elements which contributed to the success of MT implementation:

- UI strings are generally short and straightforward
- MT functionality is integrated with translation tools
- Pre/post-MT conversion process made available:
  - Source English capitalization handling
  - Multi-line handling
  - Variable handling
  - Leading and trailing space handling
  - Accelerator key handling

5 Success story of Moses

Moses was implemented as a result of an innovative project initiated by the engineering team in India. Engineers in India collaborated with the Brazilian linguist team in an effort to test Moses, an open source SBMT engine. The result was quickly evaluated and the system was moved into production within one month. The system is currently being maintained by the engineering team in India.

In addition, the Italian quality produced by Lucy did not meet expectations, compared with other language pairs. The team decided to also test Moses and the result was positive. Currently, both Moses and Lucy LT are being used in production for Italian.

6 Results

The cost savings include improved internal translation productivity, as well as the vendor cost. Comparing the standard of 2000 words per day throughput, the post-editing throughput for UI and documentation has increased as follows:

- Japanese
  - Quality stabilized, minimum update required
  - Average throughput 3000 words/day
  - Total volume processed – 4.45M new words to date
- S. Chinese
  - Quality stabilized
  - Routine maintenance on dictionary and rules
  - Average throughput 3500 words/day
  - Total volume processed – 830K new words to date
- French
  - Quality stabilized
  - Routine maintenance on dictionary and rules
  - Average throughput 3500 words/day
  - Total volume processed – 1.25M new words to date
- Italian
  - Quality not mature with Lucy LT
  - Moses is an alternative
  - Average throughput 3000 words/day
  - Total volume processed – 484K new words to date
- German
  - Quality stabilized
  - Routine maintenance on dictionary and rules
  - Average throughput 3500 words/day
  - Total volume processed – 1.22M new words to date
- Spanish
  - Quality stabilized
  - Routine maintenance on dictionary and rules
  - Average throughput 3500 words/day
  - Total volume processed – 756K new words to date
- B. Portuguese
  - Quality stabilized
  - SW quality better than documentation
  - Average throughput 3400 words/day
  - Total volume processed – 1.65M new words to date

7 MT development continues

In an effort to further improve the MT output, engineering team continues to fine tune the use of the MT engines.

- Hybrid model - combines RBMT and SBMT with a two-step processes. First translate from English to target language with RBMT engine, then process the output through SBMT against the target-target language corpus. This model is currently being tested.

- TAUS TDA data – utilize the large volume of TDA database to enhance the SBMT leveraging. Benchmark testing is in progress on Italian.

- Advanced leveraging – leverage TM at a sub segmentation level to increase fuzzy matches.

- Collaborative effort with development and tech pubs on improving English content to make it MT friendly.