Deploying novel MT technology to raise the bar for quality: A review of key advantages and challenges

Johann Roturier
Symantec
Ballycoolin Business Park
Blanchardstown, Dublin 15, Ireland
Johann_roturier@symantec.com

Abstract

This paper is a case study of the deployment of new MT technology aiming at improving the overall Post-Editing (PE) experience. One of the main challenges in having MT output post-edited within a localization workflow is being able to meet ever increasing quality expectations, especially from a terminology perspective. In Symantec's localization process we constantly seek to overcome this obstacle by refining our MT output in a number of ways. Several layers of customization have therefore been added over the years to our commercial MT system (such as User Dictionaries, Translation Stylesheets and Automated Post-Editing). Despite obtaining substantial quality gains with these techniques, improvements are still sought to minimize PE effort. The deployment of a novel technology, based on SYSTRAN's new hybrid approach, is presented in this paper. This paper focuses on the technical and linguistic challenges associated with the integration of this new technology into an existing MT workflow.

1 Introduction

In Symantec's Localization department we have been using MT in conjunction with TM technology to translate product documentation in some of our localization workflows for several years. This approach has been mostly based on an in-house optimization of a customized SYSTRAN MT system and on the post-editing of MT output by external vendors. Using MT has allowed us to increase the speed and consistency of translation, which in turn also enabled us to process larger document sets within shorter turnaround times. This paper is divided into three main sections. Section 2 focuses on the criteria we have been using for evaluating and deploying MT technology. Section 3 focuses on the customization techniques that we have been using to improve the quality of our MT output. Finally, section 4 presents the findings of a recent pilot project that was conducted to assess the deployment of a novel MT component, based on SYSTRAN's new hybrid approach. The advantages of this component will be demonstrated and the challenges encountered during the deployment of this new technology will also be discussed.

2 Defining criteria for evaluating and deploying MT technology

During our initial review of MT systems, we identified a number of criteria to evaluate and deploy MT technology. When new MT components become available, such as the one based on SYSTRAN's hybrid approach, we re-use these criteria.

2.1 Identifying the right team

While it is often suggested that (S)MT systems do not require linguists to be built, the same statement does not apply when it comes to deploying and maintaining a production MT system, especially when this system needs to be updated over-time. Evaluating the improvements achieved by a given MT system can be done semi-automatically but human checks are necessary and even essential to make sure that MT resources do not become obsolete. Deploying our MT technology took some time and this would not have been possible without the help of multiple team members, as shown by the following timeline:

- Research investigations started in 2003 (desktop products)
- Initial usage for Technical Support translations in 2005
- Enterprise system (v5) installed in 2006
- User documentation production in 2006 (EMEA) and in 2008 (APJ)
- Enterprise system (v6) deployed in 2009

Progress was slow in certain areas but the role played by our linguists (who act as MT resource owners) should not be underestimated. Without a regular maintenance of MT and TM assets, MT technology could not be used as effectively.

2.2 Identifying the right workflow

Right from the outset we decided to use MT technology (SYSTRAN 5) in conjunction with Translation Memory technology. After monitoring initial MT usage in-house, we established that using an 85% TM leverage threshold would allow us to use MT effectively by machine-translating TMX files containing unknown segments. Using this strategy allowed us to reduce translation turnarounds significantly by using in-house staff to optimize our MT resources before sending TMs for post-editing to external vendors. As an example the first large enterprise product (500K words) we localised using MT in 2007 shipped in 7 days, whereas the previous version of the same product had taken 15 days to ship. We also found that translation consistency was improved with this strategy, by having fewer bug fixes to implement in the Help system of this product.

2.3 Identifying the right evaluation metrics and tools

Traditional MT evaluation metrics such as fluency and accuracy are often measured with graded scales, which may impact the objectivity of the results. The other drawback of using these two metrics separately is that it extends the time required for the evaluation process, which in turn, increases the cost of the operation. For instance, when using two separate metrics, Coughlin (2003: 64) mentions that one has ‘to determine the importance of one characteristic over another when deciding what acceptable quality is’. In her study, which focused on the correlation between automated and human assessment of Machine Translation quality, she asked evaluators to use a unique scale of 4 values to measure the acceptability of the output. This simple approach integrated criteria concerning both intelligibility and accuracy characteristics, but was easier to use and process than if the two criteria had been evaluated separately. Based on this finding, as well as on previous research (Roturier, 2006), we designed the following post-editing specific evaluation criteria to evaluate MT segments internally:

Score	Criteria
Excellent MT output (E)	Your understanding is not improved by the reading of the ST because it is syntactically correct; it uses proper terminology; the translation conveys information accurately; minimum style requirements for Doc & Help or software content comply with the MT post-editing guidelines. **Effect:** No post-editing required.
Good MT output (G)	Your understanding is not improved by the reading of the ST even though the MT segment contains minor errors affecting any of these: grammatical (article, preposition), syntax (word order), punctuation, word formation (verb endings, number agreement), unacceptable style. An end-user who does not have access to the source text could possibly understand the MT segment. **Effect:** Only minor post-editing required in terms of actual changes or time spent post-editing.
Medium MT output (M)	Your understanding is improved by the reading of the ST, due to significant errors in the MT segment (textual coherence/ textual pragmatics/ word formation/ morphology). You would have to re-read the ST a few times to correct these errors in the MT segment. An end-user who does not have access to the source text could only get the gist of the MT segment. **Effect:** Severe post-editing is required or maybe just minor post-editing after spending too much time trying to understand the intended meaning and where the errors are.

Poor MT output (P)	Your understanding only derives from the reading of the ST, as you could not understand the MT segment. It contained serious errors in any of the categories listed above, including wrong Parts Of Speech. You could only produce a translation by dismissing most of the MT segment and/or re-translating from scratch. An end-user who does not have access to the source text would not be able to understand the MT segment at all. **Effect:** It would be better to manually retranslate from scratch (post-editing is not worthwhile).

Table 1. PE-specific MT evaluation criteria

When post-edited content is returned by our vendors, our team of linguists performs an evaluation on sample segments using the criteria defined in the table above. This allows us to make sure that the performance of our MT system maintains specific quality levels. We also supplement human evaluation with automatic scores.

Finding an automated metric that correlates well with human scores can be a challenge. BLEU is often used to report on system improvements but its correlation with human scores is disputed, especially when it comes to evaluating a rules-based MT system. After reviewing a number of metrics internally, it was found that GTM (Turian et al. 2003) correlated well the human evaluation criteria described earlier (when used with a 1.2 exponent). Correlation results are presented in Section 4.5.

While GTM scores (or any other automatic score for that matter) obtained at a segment level can be a useful indicator, using the average of all scores in a project can be misleading if the length of source segments is not taken into account. Obviously the PE effort is not going to be same for a short segment (say under 6 words) than for a long segment (over 20 words). So if the average score of all segments were to be used, the wrong picture may emerge. To address this issue, we have come up with a custom way to calculate a "project score" when evaluating a set of files. This project is calculated in the following manner:

- Eleven score categories are defined for every increment between 0 and 1 (such as >=0<0.1, >=01<0.2, etc.)
- Each score category is given a coefficient that is used as a multiplier (for instance the coefficient for the category >0<0.1 has a coefficient of 0, the coefficient for the category >=01<0.2 a coefficient of 0.15, etc.)
- Each segment is given a GTM score and the number of words present in the source segment is added to a score category pool (as defined in the step 1)
- The percentage of each score category pool is then calculated based on the overall number of words in the project.
- The value of each score category is calculated by using the coefficient defined in step 2
- A final project is calculated by adding all of the score categories' contributions.

This technique rewards long segments and makes sure that short and "easy" segments do not contaminate the overall project score.

2.4 Identifying appropriate source content

The following quality requirements exist for the translation of product documentation:

- Consistent Key Terminology (which can be enforced by using an MT dictionary)
- Focus on Features and Product Names (which can be enforced through terminology Preparation and machine-translation)
- Correct Software References (which can be enforced by using a specific user dictionary)
- Translated content correctly reflects source content and has no negative impact on comprehension

We feel most of these requirements can be met by relying on quality MT user dictionaries. However, the quality of these user dictionaries will not compensate for uncontrolled source content. We found that source control was a pre-requisite to MT success so we decided to invest time in transforming existing editing guidelines into Controlled Language rules. Since 2006 some of our writing teams have therefore been using a Controlled Language checker (acrolinx™ IQ suite) during the authoring process to make sure that source content complies with pre-defined spelling, terminology, grammar and style rules. We have shown in previous reports (Roturier, 2009) that the lower the source project score (in terms of violations of style, grammar and spelling), the higher the MT scores (be they human scores or automatic scores).

3 Review of existing MT customization techniques

Several customization techniques are used to ensure that the MT output that is sent for post-editing reaches a quality level to makes post-editing worthwhile. These techniques involve custom MT user dictionaries, custom translation stylesheets and an automated post-editing module.

3.1 Custom MT User Dictionaries

A list of new terms is compiled before each project and preferred translations are identified by our in-house linguists. This list of terms is then imported into Symantec's user dictionaries and coded using Systran's Intuitive Coding technology (Senellart et al., 2001). The main advantage of this approach is the speed at which terms are imported and coded, where other MT systems may require additional linguistic information for all new terms. Certain entries may generate problems during the coding process, but these can be easily rectified by adding special coding clues (Ibid) to ensure that the system uses the entry properly. The following example demonstrates the use of a coding clue (between brackets) for an English to German entry: *to point to > zeigen auf (governs_accusative)*

While it can be argued that the time spent creating User Dictionary entries could be spent translating documents, the benefits of building quality terminology resources should not be underestimated since these User Dictionaries can be re-used in other contexts (for example for the translation of technical support documents).

3.2 Translation Stylesheets

SYSTRAN's stylesheet technology (Senellart and Senellart, 2005) uses XSLT to drive and control the machine translation of XML documents and allows for context-sensitive translations. Without them, input files (such as TMX or XLIFF files) that are generated after the TM analysis step could not be translated. We decided to use custom stylesheets for TMX files in order to handle our native XML format (a subset of Docbook which is embedded in TMX or XLIFF files) more effectively. For instance XSL rules were written to separate the translation of GUI options (marked with *guimenuitem* or *guilabel* tag) from the translation of the other parts of a segment. Language-specific rules were also written, for instance to have an English string next to a Simplified Chinese string for every GUI option present in our documentation. Since this type of requirement varies from project to project, we have made several translation stylesheets available to our project coordinators who are in charge of file processing. This approach, which is extremely flexible, also allows for making sure that non-translatable content (contained in *command* or *userinput* elements) does not get translated: very often, it is indeed difficult to prevent an MT system from over-translating.

3.3 Automated Post-Editing

As mentioned by O'Brien (2002), translators are often reluctant to post-edit machine translation output because of their 'dislike for correcting repetitive errors that a human translator would never make'. To fix repetitive errors that could not be fixed with SYSTRAN's normalization dictionaries, a custom post-processing approach is used to automate the post-editing task. The concept of Automated Post-Editing was first introduced by Knight and Chander (1994) and further explored by Allen and Hogan with a view to fix 'systematic errors committed by an MT system' (2000). When these MT errors cannot be fixed with advanced User Dictionary coding techniques, they may be fixed using powerful global search and replace patterns. Our post-processing module is based on regular expressions (Roturier et al. 2005). For instance the following example shows that a local word order problem (position of filenames) is a fixed in generic manner:

- MT output before post-processing: Die Local.cfg Datei unterstützt nicht Sprachen, die DoppeltByte Zeichensätze verwenden.
- Search pattern: (?<!\boder)(\b\w+?\.\w{1,5}) (Datei\b)
- Replace pattern: $2 $1
- MT output after post-processing: Die Datei Local.cfg unterstützt nicht Sprachen, die DoppeltByte Zeichensätze verwenden.

While the advantages of such an approach are clear, Allen and Hogan (2000) envisioned that an 'APE system might hypercorrect words and grammatical structures that are acceptable in the MT output'. The impact of this approach has been previously reported by Roturier and Senellart (2008), whereby extremely high improvement/degradation ratios were obtained (in the region of 50) while showing modest automatic score improvements (+1 GTM point).

4 Deploying a novel MT component

4.1 Review of previous work

A number of hybrid experiments have recently been conducted by combining rule-based MT (RBMT) systems with Statistical Post-Editing (SPE) systems. Two experiments were carried out for the shared task of the ACL 2007 Workshop on Statistical Machine Translation, combining a raw SYSTRAN system with a statistical post-editing (SPE) system. One experiment was run by NRC using the language pair English<>French in the context of 'Automatic Post-Edition' systems using the PORTAGE system as described in Simard et al. (2007). The second experiment based on the same principle was run on the German > English and Spanish > English language pairs using the Moses system (Koehn et al. 2007). The objective was to train a SMT system on a parallel corpus composed of SYSTRAN translations with the referenced source aligned with its referenced translation.

A detailed evaluation of these experiments was then conducted and presented in Dugast et al. (2007). They concluded that the SYSTRAN+SPE experiments demonstrated very good results – both on automatic scoring and on linguistic analysis. Their detailed comparative analysis provided directions on how to further improve these results by adding "linguistic control" mechanisms. Finally, their results also set a baseline to compare with other more sophisticated/integrated "rules and statistics" combination models. The pilot project presented in Section 4 builds on these experiments, by expanding the number of language pairs being used, as well as by diversifying the validation and evaluation metrics.

4.2 Objectives

The main objective of this pilot project was to improve quality and the overall post-editing experience by supplementing our existing customization components with SPE models.

We also wanted to make sure that the following linguistic features would be supported:

- Recase output
- Preserve key terminology
- Support for tagged input
- Show positive Improvement/Degradation ratio for all language pairs (using error weights)

From a performance perspective, it was also hoped that the SPE models would not increase our baseline processing time by more than 30% (which currently reaches throughput rates of around 200 tokens/second for the translation of TMX files).

4.3 Pilot project setup and validation

In May 2009 we decided to conduct a pilot project to evaluate the effectiveness of SYSTRAN's new hybrid component. Since the training of SPE models was not supported in the version 6 of SYSTRAN Enterprise Server, we had to send our TMs to SYSTRAN with a view to use the resulting SPE models for a large production project.

The first challenge was to make sure that the TMs sent would be as clean as possible (to avoid noise during the training phase). However, we found out that TM cleanup is a task that is poorly documented in the industry and not well supported by traditional translation tools. We therefore relied on internal tools to check for key translation consistency.

We also had to make sure that we kept a set of evaluation segments to validate the new component. Keeping these two requirements in mind, sets of 40K translation units were sent for off-line training to SYSTRAN. The resulting SPE models were then delivered and deployed in a test environment. The evaluation set was then used to validate quality gains using the project scores described in Section 2.3. These scores are presented in Table 2.

Language /Score	Systran 6.06 (with default Symantec UDs)	Systran Hybrid (with default Symantec UDs)
Italian	46.42	56.24
French	51.91	56.72
Japanese	45.94	58.88
Simp. Chinese	52.67	58.14
German	37.67	45.94

Table 2: Project scores obtained during validation

4.4 Challenges

One of the main challenges we encountered when deploying the SPE models was to find a way to visualize the differences brought by the SPE models. This was addressed by creating an internal tool that acts as a wrapper for a string diff library. The output is then generated in Excel so that linguists can check where improvements or degradations occur.
Another challenge was to find a way to effectively integrate the SPE models into our existing MT/TM workflow. This was due to the fact that TM cleaning is still an area that is under-researched as no standard tool exists to safely perform global terminology replacements, especially for languages that are highly inflected.
Besides, having to rely on an off-line process to generate the SPE models was a challenge. This meant that the User Dictionaries that were used to train the SPE component drifted slightly from the actual SPE components.
Finally performance proved to be a minor issue as the processing time took longer than expected (around 5 times longer than our normal processing times). While this would have a minor impact on our translation workflow, it is still acceptable for our use of MT, which attempts to deliver the best output possible to post-editors.

4.5 Results

In this project the effectiveness of the new approach is evaluated by focusing on technical post-editing effort (following Krings' division of post-editing effort into cognitive, technical and temporal (2001). While technical post-editing effort can be measured by using sophisticated techniques and tools (O'Brien, 2006), the approach used in this paper is to examine how close an MT segment is from the reference translation.
The following results were also obtained after sampling a number of segments from the validation set and performing a manual error classification analysis. We then calculated the number of improvement/degradation ratios in various error categories. The error category classification is mostly based on Dugast et al. (2007), to which one extra category was added (tags). To obtain the ratios, the number of improvements is divided by the number of degradations. When divisions do not produce valid results, scores are indicated with a * (no degradation) or a hyphen (no improvement, no degradation.)

Category/Lang.	JA	CS	FR	IT
termchg_nfw*	0.7	2.8	-	1
termchg_term	5.3	3.9	2	5
termchg_loc	2	0	1	7*
termchg_mean*	0.54	1.23	1	0
gram_det	-	5	7	0.5
gram_prep*	1.34	1.85	5.75	6
gram_pron	-	0.5	1*	8*
gram_tense	1	1*	7	10
gram_number	1*	-	1	0.5
gram_gender	-	-	1	-
gram_other	2.75	-	3	26*
punct/digit/case	0.2	-	0	1*
wordorder_short	0.17	-	1*	0.7
wordorder_long*	1	-	-	3
tags*	1*	-	-	-
Average	1.45	2.2	3	3

Table 3: Error category analysis

This table shows that improvements are not always consistent across language pairs. For instance, changes affecting meaning (termchg_mean) were more prevalent in Japanese than in the other languages. On the other hand two categories show clear improvements (termchg_term and gram_prep) which suggest improvements in terms of fluency. Slight degradations are observed in the handling of punctuation, digit and case, showing that the core SYSTRAN engine is better equipped to deal with these linguistic features than the SPE component. However, tags did not suffer from any degradation because the handling of tags, which is performed by our custom stylesheet (as described in Section 3.2), allows for the linking of target tokens with tags, which means that the SPE component is able to preserve tag position in the target output.

Finally, we also sent 5K words for post-editing in four language pairs, after machine-translated this content using the SPE models delivered by SYSTRAN. It should be noted that this document set was in beta form as it had not been through a full source checking cycle, so we were aware that this would create additional problems during the MT process. The objective of this pilot project was to get feedback from post-editors and check whether this feedback would correlate with the score improvements observed during the validation step. Based on this analysis, we hoped to make a decision on the use of these SPE models on a larger document set as part of a production project. The following project scores were obtained:

Language /Project Score	SYSTRAN Hybrid (with default Symantec UDs)
Italian	66.86
French	62.97
Japanese	70.34
Simp. Chinese	61.66

Table 4: Project scores obtained during the pilot project

These scores were supplemented by an in-house analysis of the MT segments using the evaluation criteria defined in Section 2.3. For French and Italian, the scores obtained respectively were: Excellent (48% and 38%), Good (26% and 27%), Medium (9% and 22%) and Poor (20% and 13%). When calculating the Pearson correlation between these scores and the GTM scores at the segment level, high correlation is achieved (0.86 for Italian and 0.76 for French).

Feedback was also obtained from post-editors. Feedback varied somewhat from one language pair to the next. For instance, for French and Italian, our vendors found that "throughput was improved slightly." But they also noted that the files contained a lower number of complex sentences, which may be attributed to our efforts controlling source content. So they concluded by stating that "the overall experience was a little better". This is confirmed by the feedback obtained for Chinese, where the vendor noticed "progress in fluency and meaning", but did not notice "big improvement in other aspects, especially in proper translation and punctuation". Overall, they felt their PE experience improved by 8% by comparing the amount of time they spent post-editing these segments (compared to the amount of time they had spent post-editing standard output). On the other hand, the PE experience for Japanese was not as positive. The vendor felt that the quality was worse, and that "most of the sentences had to be rewritten to completely different sentences" (even though 203 segments out of 480 did not require any change during the PE step). Specifically they highlighted the fact that "words in the original sentence were not in output sentence" (for example, important negative words such as "not" disappeared) and that "words that were not in original sentence were used in output". This suggests that the statistical layer affected the meaning of certain translations, which in turn had a negative impact on the original confidence associated with the rules used by the core SYSTRAN engine. This certainly requires further research. Another interesting finding is that in some cases, the vendor found that some "liberal translation departing from original sentence was re-used". This suggests that the TM cleanup step must be better understood to possibly filter out sentences that should not be re-used during the training process.

5 Conclusions and next steps

This paper has shown that a novel MT component could be deployed to improve the overall PE experience in certain language pairs. While some advantages have already been noticed (such as improved fluency), some challenges still have to be addressed. The three main challenges are:

- Reducing degradations to a minimum (if this is not possible, then relying on a pure rules-based output may be preferable for certain types of sentences or if sentences contain or do not contain certain constructs). This decision should be made by the MT engine based on its confidence of the MT outputs it can produce.
- New terminology still needs to be identified, defined and encoded (whether in a User Dictionary or in a post-editing string). Clearly this step should be further automated, possibly using a collaborative ap-

proach to suggest and validate new terminology.
- TM cleanup and management should be further investigated to isolate segments that are not worth re-using at a sub-segment level. Selecting quality data rather than large data sets seems preferable when the objective is to improve PE experience.

Acknowledgments

I would like to thank my colleagues at Symantec without whom this recent work would not have been possible: Fred Hollowood, Orla Clifford, Katrin Drescher, Nathalie Moyano, Laure Jouanique, Flora Iacoponi, Julia Schinharl, Liliane Seifert, Jun Sakai and Zoe Chen. I would also like to thank Dr. Jean Senellart and Dr. Elsa Zipstein-Sklavounou for sharing their expertise when some of the evaluation work was conducted. Finally I would also like to thank Nora Aranberri for providing comments on an earlier version of this paper.

References

Chris Callison-Burch, Miles Osborne, and Philipp Koehn, 2006. Re-evaluating the Role of Bleu in Machine Translation Research. In Proceedings of EACL-2006.

Deborah Coughlin. 2003. Correlating Automated and Human Assessments of Machine Translation Quality. In Proceedings of MT Summit IX, New Orleans, USA. pp. 63-70.

Hans. P. Krings. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Postediting Processes*, G. S. Koby (ed.), Kent, OH: Kent State University Press.

Jean Senellart, Jin Yang, & Anabel Rebollo. 2003. SYSTRAN intuitive coding technology. MT Summit IX, New Orleans, USA, 23-27 September 2003; pp.346-353.

Jeff Allen and Christopher Hogan. 2000. Toward the Development of a Post-Editing Module for Raw Machine Translation Output: A Controlled Language Perspective. In Proceedings of the Third International Workshop on Controlled Language Applications, Seattle, WA, pp. 62–71.

Johann Roturier. 2009. Controlled Language for MT in Action. Presentation given at Translingual Europe 2009, Prague. Available at http://ufal.mff.cuni.cz/tle2009/presentations/roturier-controlled-language-for-mt-in-action.pptx

Johann Roturier and Jean Senellart. 2008. Automatic Post-Editing: Review of Translation Quality Gains. Presentation given at LISA Forum 2008, Dublin.

Johann Roturier. 2006. An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness, and Acceptability of Machine-Translated Technical Documentation for French and German Users. Unpublished PhD thesis, Dublin City University, Ireland.

Johann Roturier, Sylke Krämer, and Heidi Düchting. 2005. Machine Translation: The translator’s choice. In Proceedings of the 10th LRC conference, Limerick, Ireland.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of Machine Translation and Its Evaluation. Proceedings of MT Summit 2003:386-393. New Orleans, Louisiana.

Kevin Knight and Ishwar Chander. 1994. Automated Post-Editing of Documents. In Proceedings of the 12th National Conference on Artificial Intelligence, Seattle, WA, pp. 779–784.

Loic Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical Post-Editing on SYSTRAN’s Rule-Based Translation System. In Proceedings of WMT-2007. Pp. 220-223.

Michel Simard, Nicola Ueffing, Pierre Isabelle and Roland Kuhn. 2007. Rule-based Translation With Statistical Phrase-based Post-editing. In Proceedings of WMT07. Pp. 204-206.

Pierre Senellart and Jean Senellart. 2005. SYSTRAN Translation Stylesheets: Machine Translation driven by XSLT. In Proc. XML Conference & Exposition, Atlanta, USA, November 2005.

Sharon O’Brien. 2002. Teaching Post-Editing: A Proposal for Course Content. In 6th EAMT Workshop Teaching Machine Translation, Manchester, pp. 99–106.

Sharon O'Brien. 2006. Methodologies for Measuring the Correlations between Post-Editing Effort and Machine Text Translatability. In *Machine Translation*. Vol. 19, No. 1. pp. 37-58.