A Comparison of Statistical Post-Editing on Chinese and Japanese

Midori Tatsumi **Yanli Sun**
School of Applied Languages and Intercultural Studies
Dublin City University
Dublin 9, Ireland
midori.tatsumi2@mail.dcu.ie yanli.sun2@mail.dcu.ie

Abstract - *This paper analyses both quantitatively and qualitatively the results of a recent Statistical Post-editing (SPE) experiment on English to Chinese and English to Japanese translations. Quantitatively, it compares the number of changes resulting from SPE between the two languages; qualitatively, a linguistic analysis of the changes is conducted. It also investigates the effect of SPE on the fluency and adequacy of the translation as well as the potential impact on human post-editing effort. Our study indicates that, in general SPE results in more improvements than degradations in both languages although the linguistic changes are different between the two languages. In addition, SPE could improve the fluency and adequacy of MT outputs and shorten human post-editing time in both languages.*

Keywords: Statistical Post-Editing, RBMT, SMT, Chinese, Japanese

1 Introduction

None of the Machine Translation (MT) systems that are currently available are good enough to produce error-free outputs, and as Allen & Hogan (2000) point out, MT errors are likely to recur throughout or across documents. Therefore, post-editors are often dispirited by the need to make the same correction over and over again (Isabelle et al 2007: p 255). In order to ease the burden placed on human post-editors, Allen & Hogan (ibid) proposed the development of an automatic post-editing (APE) module that would automatically repair mistakes in raw MT output by utilising the information on the changes that were made during the post-editing process from “parallel tri-text (source texts, MT output, post edited texts)” (Allen & Hogan 2000: p 62). Elming (2006) presented the first results of the use of an APE module to correct the output of a rule-based machine translation (RBMT) system and it was noted that translation quality increased noticeably in terms of BLEU scores (an automatic machine translation evaluation metric) (Papineni et al 2002).

The advent of statistical machine translation (SMT) not only presented an entirely new method of machine translation, but also opened the door to the possibilities of combining two different MT systems to benefit from the advantages of both. Knight & Chander (1994) proposed to use SMT techniques to learn the mapping between a large corpus of “pre-postedited” (ibid, p 779) texts with aligned corresponding post-edited text. Simard et al. (2007a, 2007b) tested and extended this proposal by using a statistical phrase-based MT system to post-edit the output of an RBMT system. The basic mechanism of this kind of system, which is now often referred to as a statistical post-editing (SPE) module, is as follows: an SMT system is trained using a set of raw RBMT output and its corresponding reference text, which is either human post-edited or human translated (training corpora). In this way, SMT learns how to “translate” raw RBMT output to better quality text. Their experiments showed that this SPE module could improve the quality of the RBMT output. However, a detailed analysis of the improvements and degradations of SPE in the previous experiments had not been presented until Dugast et al. (2007) described their experiment on a combination of Systran and SPE. They evaluated, qualitatively, the changes made by SPE modules on the output of Systran, including some linguistic analysis such as improvements, degradations and equivalent effects. However, as with most of the previous studies, their study was only conducted on European language pairs. Until recently, little research has been done on the effect of SPE on Asian languages such as Chinese and Japanese.

One such instance of this research is an experiment conducted in 2008 by Systran and Symantec (Senellart & Roturier forthcoming) to investigate the potential of SPE when used in combination with Systran, an RBMT system. The general procedure of the SPE process used in the experiment was as follows: Systran 5.05 was used as the RBMT system, and Moses (an open-source toolkit for statistical machine translation) (Dugast et al 2007, Koehn 2004) was used as the SPE tool. All of the training and test resources were provided by Symantec, which included translation memories (TM) and user dictionaries (UD), in the following language pairs: English to French, German, Chinese, and Japanese. Four parallel corpora have been produced for each language: translation by Systran without UD (referred to as *Systran – Raw* for the purposes of this paper), Systran translation with UD (*Systran - Customised*), Systran translation without UD, combined with SPE (*Systran - Raw & SPE*), and Systran with UD, combined with SPE (*Systran – Customised & SPE*).

The current paper analyses and compares, both quantitatively and qualitatively, the Japanese and Chinese output of *Systran – Customized* and *Systran – Customized & SPE*. The experimental setting for these two languages is as follows: Chinese (ZH) TM consisted of 529,822 translation units of English source texts and corresponding target text, while Japanese (JA) TM consisted of 143,742 units. Both TMs were created based on human translation, instead of human post-edited MT output due to insufficient post-edited data. The UD included Symantec-specific user interface terms as well as general terms to which certain target language words had been assigned. The UD contained 8,832 entries for Chinese and 6,363 for Japanese.

A preliminary classification and evaluation of the changes made by SPE on Chinese and Japanese is conducted in Section 2. Sentence level human evaluation results are presented in Section 3. Finally, Section 4 concludes this study and points out future work.

2 Classification and Evaluation of Changes

2.1 Evaluation setup

As mentioned earlier, in this study, the authors (one Japanese and one Chinese) have decided to carry out detailed linguistic comparisons of the results from *Systran - Customised* and *Systran - Customised & SPE*, since *Systran - Customised* is the standard MT translation method currently employed by Symantec, and we are interested in what would happen if the SPE process was added to the current standard operating procedure. The aforementioned experiment, by Senellart & Roturier (ibid), has shown that *Systran - Customised & SPE* outperformed *Systran - Customised* for both Chinese and Japanese in terms of BLEU and GTM (Turian et al 2003) scores. The BLEU score rose by about 6 points and 12 points, and the GTM score by about 10 points and 7 points for Chinese and Japanese respectively. To reveal the detailed linguistic changes that have caused these improvements in performance, the authors randomly selected a sample of 100 translation segments from each of the Chinese and Japanese test sets, and conducted a quantitative and a qualitative evaluation of the results, comparing the source text, Systran output, SPE output, and the reference human translation to see how many and what types of improvements and degradations had been made during the SPE process.

The quantitative evaluation was conducted using evaluation categories defined by the authors based on the Error Classification suggested by Vilar et al. (2006). The classification was modified to make it more suitable for categorising changes rather than errors, and simplified to ensure applicability to both Chinese and Japanese. The changes made during the SPE process were categorised into Words/Phrases Alteration/Deletion/Addition, Forms (Tense or Voice, Formality, and Imperative), Translation of Fixed Expression, Word or Phrase Reordering, and Punctuation. The number of improvements, degradations, and equivalent changes in each category was counted. It was decided to adhere strictly to the reference translation when assessing each change in order to avoid the subjectivity of the authors and standardise the evaluation process. Therefore, when the translation of a word in either MT output or SPE output did not match with the one in reference translation, it was regarded as an “equivalent change” even if the change made during the SPE process seemed to have improved the quality of the translation.

The qualitative analysis was performed after the quantitative evaluation in an effort to explain some of the most common changes made during the SPE process. Firstly, similarities and differences between the language pairs were identified, and the factors responsible for these similarities and differences were studied.

2.2 Results and discussion

Below is the result of the quantitative evaluation of the changes that resulted in improvements, degradations, and equivalent effects for Chinese (ZH) and Japanese (JA) respectively.

Change Categories		Improvement		Degradation		Equivalent	
		ZH	JA	ZH	JA	ZH	JA
Word/Phrase Alteration	Content Words	137	45	19	40	28	25
	Function Words	38	45	6	9	17	30
Word/Phrase Deletion	Content Words	0	9	0	2	0	1
	Function Words	51	57	4	5	12	16
Word/Phrase Addition	Content Words	4	0	3	2	2	0
	Function Words	12	1	8	2	15	1
Forms	Tense or Voice	6	3	0	0	3	5
	Formality	0	1	1	0	0	0
	Imperative	0	8	0	0	0	2
Fixed Expression		8	0	0	0	0	1
Word/Phrase Reordering		9	1	3	3	0	1
Punctuation		31	47	4	9	0	4
Total		296	217	48	72	77	85

Table 1. Number of Improvements, Degradations and Equivalents in ZH and JA

As can be seen from the table, the number of improvements for Chinese text is noticeably higher than Japanese, and the number of degradations for Japanese is noticeably higher than Chinese. Based on the evaluation that we conducted in this research, it can clearly be seen that the SPE process has had a more positive impact on the Chinese text than on the Japanese text.

Most notably, there have been numerous improvements in the choices of content words/phrases for Chinese, which happened three times more often than in the Japanese text. For Japanese, the changes that were made to content words/phrases have done as much harm as good. Another thing that is worth mentioning is the changes that were made to punctuation, which have had an equally beneficial impact on both Chinese and Japanese.

Based on the quantitative analysis, we find that the most frequently changed categories are similar in Japanese and Chinese, be they improvements or degradations, such as function words/phrases alteration and function words/phrases deletion. On the other hand, there are also great differences between the two languages, for instance, there have been no content word deletions in Chinese while there have been some in Japanese. A detailed investigation on what constitutes those changes and whether the same category contains the same linguistic changes in Japanese and Chinese has also been conducted.

2.2.1 Similar effects of SPE between the two languages

By “similar effects”, we mean those categories that share almost the same level of changes after SPE in Japanese and Chinese. SPE had a similar effect on the following categories in Japanese and Chinese: function words/phrases (alteration or deletion) and punctuation.

Improvements in function words/phrases alteration

One of the most prominent similarities observed in this study is found in the changes made by SPE on function words/phrases. The Improvement/Degradation rates for the Alteration of Function Words were 6.3:1 and 5:1 for Chinese and Japanese respectively, and the rates for the Deletion of Function Words/Phrases were 12.7:1 and 11.4:1 respectively.

Some of the function words/phrases alterations have been made in a very similar manner for both Japanese and Chinese. One example of this would be the changes to more appropriate translations for certain prepositions, such as “to” and “about”. Another example of common types of function words/phrases alteration is the correction of modal verb translations such as “can” or “must”. In Table 2 below, MT output refers to the output of *Systran – Customised* while SPE output refers to the output of *Systran – Customised & SPE* as we mentioned in section 1. Glosses are omitted due to the fact that SPE mostly makes subtle changes by using more appropriate or desired words in the specific context, and the basic meaning often remains the same.

Source	MT output	SPE output
To maintain …	JA: 保守するため…	維持するには…
Reverts to	ZH: 恢 复 对	恢 复 到
must configure	JA: 設定しなければなりません	設定する必要があります
You can …	ZH: 您 能	您 可 以

Table 2. Example of Function Words/Phrases Alteration

However, within the same categories, there have also been some differences in the types of alteration, presumably mostly due to the language differences. A couple of examples for Japanese cases are shown in the table 3. The first one is a stylistic change of character types from Kanji (ideograms) to Hiragana (phonetic characters), which could also be handled by a simple global search and replace operation in any text editor. However, the second one may be a good example of SPE-specific abilities, where the subjective postposition has been changed to one that is more appropriate in the specific context.

Source	MT output	SPE output
(Imperative sentence ending)	JA: して下さい	してください
Messages are deleted	JA:メッセージは削除されます	メッセージが削除されます

Table 3. Example of unique alteration in JA

Specific changes in Chinese include translations for some of the relative pronouns, demonstrative pronouns, and quantifiers being changed to more appropriate ones during the SPE process. For example, the translation of “this” was changed from “此” to “该”. Although these two Chinese characters share the same meaning and their back translation is probably the same, the second one is the more commonly used word in the reference translations.

Improvements in function words/phrases deletion

One common improvement as a result of the Deletion of Function Words/Phrases among Chinese and Japanese was the desirable omission of personal pronouns, such as “you” and “they”, which are commonly dropped both in Chinese and Japanese. Yoshimi (2001) has suggested a method of eliminating or substituting unwanted pronouns in English to Japanese machine translation without human intervention using a decision-tree learning method. In his method, however, the corpora for statistical learning must be created by a human for generic purposes, whereas the training corpora in the current research have been compiled automatically from the very specific domain text, and have been proven to be effective. In Table 4, the underlined translation was omitted during SPE.

Source	MT output	SPE output
the actions that you specify for that rule	JA: あなたがその規則のために指定する処理	そのルールに指定する処理
After you configure your	ZH: 在您配置您的	配置

Table 4. Function Words/Phrases Deletion

Other than personal pronouns, there are no notable similarities in the types of deletions made to the two languages. For Japanese, a number of improvements were made by the positive deletion of unnecessary prepositions, such as "for" (ための), and unnecessary sentence endings caused by wrong part-of-speech parsing. For instance, "definition files" was originally translated as a sentence "定義はファイルします" [The definition files (something)], which has been correctly changed to a noun phrase "ファイル定義" [definition file] during SPE. In Chinese, the deletion of unnecessary translations for quantifiers is quite common, for example, the translation of "Provides a more detailed explanation" [MT output: 提供一个详细说明] is modified in SPE by deleting the translation of "a' [SPE output: 提供详细说明].

Improvements in punctuation

Another notable similarity is found in the changes made to punctuation. The Improvement/Degradation rates were 7.5:1 and 5.1:1 for Chinese and Japanese respectively. In the case of Japanese, one of the major reasons for improvement was the successful deletion of unnecessary hyphens that had been inserted during the RBMT process. Another major positive impact was due to the alteration of the type of full stops to ones that are preferred in the specific context of Symantec. In the case of Chinese, one improvement is the deletion of incorrectly generated commas in front of sentences as the first Chinese example in Table 5 shows. Another improvement is the correct alteration of regular commas into special Chinese enumeration commas when separating items constituting a list, see the second Chinese example in Table 5.

Source	MT output	SPE output
MPE provides an option …	JA: オプションを提供します。	オプションがあります.
Control Centre performance may be diminished while the synchronization is in progress.	ZH: ，当同步进展中时…	同步处理….
You can add, edit, copy, delete …	ZH: 您能添加，编辑，复制，删除	您可以添加、编辑、复制、删除

Table 5. Punctuation changes in JA and ZH

2.2.2 Different effects of SPE between the two languages

Difference here refers to the fact that the influence of SPE is not universal for all categories within the two languages; certain categories in one language are influenced much more than those in the other.

Improvements and degradations in content words/phrases

The most notable difference between the results of the two languages might be the changes made to content words and phrases. Caution must be exercised not to overestimate the number of improvements made by the content words/phrases alteration for Chinese since there have been a large number of repeated identical changes in the analysed Chinese segments. Nevertheless, the Improvement/Degradation rate of 7.2:1 is worth investigating especially considering the significantly low rate of 1.1:1 for Japanese.

For Chinese, there are several kinds of improvement. One improvement is the alteration of nouns, be it general terms or domain-specific terms. Firstly, terms that were originally not translated have been translated after the SPE module, for example "sub domains" which is not translated by RBMT, is translated into "子域" as in the reference translation; secondly, terms have been translated more appropriately. For example "scanner" whose translation is "扫描设备" by RBMT, was changed to "扫描器", the same as the reference translation. The appropriate adaptation of verb translation is another major reason for improvement, for example, "recommend" was translated into "推荐" while the SPE module changed it into another, more desirable, translation "建议". More appropriate choices of adverbs and adjectives are the other reasons for the improvement, such as changing the translation of "unchecked" from "未经检查的" to "未经选中的" etc.

Most of the alterations of content words/phrases have had a positive effect, however, there are cases where the changes have had a negative effect, for example, "extra" is correctly translated by RBMT as "额外的", however, SPE incorrectly changed it into "无关的[unrelated]".

For Japanese, almost the same number of improvements and degradations have occurred. One of the notable reasons for improvement was the correction of the part-of-speech parsing. For example, noun phrases, such as "filtering rules" and "console commands", which had originally been translated by RBMT as sentences, such as "フィルタは适配します [The filtering rules (something)]" and "コンソールは命じます [The console commands (something)]", were properly converted back to noun phrases in Japanese as a result of the correction of mistranslations of plural nouns as third person verbs by RBMT. Another reason for improvement was the achievement of better collocation. For example, the translations of certain words, such as "grant" or "unwanted" need to be carefully selected depending on their collocations, and some of the incorrectly selected terms in RBMT output were changed to words that were more appropriate in each circumstance. In addition, some of the translations of domain specific terms, such as "rule" and "run" have been found to be translated into more appropriate Japanese words.

However, many of the degradations also resulted from the mistranslation of domain specific terms. For example, the words "document", "store", and "alert", which had originally been translated properly, conforming to the user dictionaries, were incorrectly changed to different terms. Misinterpretation of general terms has also occurred. For example, the word "number", which had correctly been translated as "番号 [sequential number]", was changed to "数 [quantity]". In addition, there has been an instance of case confusion, where a correctly translated instance of "it" was changed to "IT (Information Technology)". Also, some of the correctly translated words have been replaced with different words; for example, "バックアップ文書(backup document)" changed to "バックアップデータ (backup data)" or "バックアップファイル (backup file)". Such degradations might be attributed to using human translation as the training data of SPE, which may have included more variations in translating the same English words than human post-edited MT text.

<u>Function words/phrases addition</u>

Another major difference was the addition of Function Words/Phrases. Since there are no inflections and derivations in Chinese, which means that Chinese is not a morphologically rich language compared to most European languages, the Chinese language uses additional function words to express tense or voice. For example, "A black dash indicates that it <u>is</u> disabled" is translated as "黑色破折号表明它禁用", the SPE correctly modifies this into "黑色线表明它<u>已</u>禁用" by adding a word expressing the tense and voice. Also, some English prepositions should be translated into circumpositions in Chinese, which require an additional character placed after the phrase. For example, "<u>on</u> the Spim tab" is originally translated into "在Spim选项卡" and later changed into "<u>在</u>Spim选项卡<u>上</u>" with the underlined word added to express the full meaning of the preposition "on".

Fixed expressions

Another difference between Chinese and Japanese is that fixed phrases in English have been translated into more appropriate Chinese phrases as in the reference translation. For example, "In general" has been modified to "通常情况下" from "一般情况下". This type of change has not happened at all in the Japanese text.

Words/phrases reordering

Word order is one of the most important factors for determining the meaning of a sentence in Chinese. Correct order is vital in adequacy and fluency. There are cases where SPE has corrected some incorrect word order, such as, "These threats are then…" is translated literally as "这些威胁然后", while SPE modified it into the correct order as in the reference translation "然后，这些威胁" by putting "then" in font of the Chinese sentence. While words/phrases reordering happened frequently, and often with a positive impact in Chinese, it occurred only a few times in the Japanese text and three occurrences resulted in ungrammatical sentences, for example sentences that begin with an adverbial particle.

Form of Imperatives

One of the interesting differences may be the changes made to the sentence pattern, which happened relatively frequently in Japanese but not at all in Chinese. Changes from the polite imperative form [して下さい] to the polite basic form [します] occurred 10 times in the Japanese text, eight of which had positive effects. This change conforms to the sentence pattern commonly preferred in user manuals.

2.2.3 Errors that have not been corrected by SPE

There have been similar errors produced by Systran in both languages outputs that were not corrected by SPE. Firstly, long range word reordering rarely happened, thus when the MT system produced improper sentence structures, mainly due to misplacement of clauses or incorrect parsing of prepositional phrase attachments, they were very rarely corrected. Secondly, intelligibility was rarely improved, even when local changes such as terms and function words were altered. These examples may suggest that it may not be appropriate to expect that sentence level correction can be achieved by SPE processes.

One error produced frequently by Systran in Chinese is the mistranslation of "and" conjunction phrases, which were not corrected during the SPE process, whereas such a mistranslation was rarely observed for Japanese. Secondly, some of the terms that should remain in English have been undesirably translated into Japanese, for example, the translation of "OLE" to "オーレ", which was not observed in the Chinese text and was not corrected in the SPE process. Finally, some of the user interface terms have been unnecessarily translated in the Japanese text and remained incorrect during the SPE process. On the other hand, in Chinese, some of the user interface terms that Systran failed to translate were successfully translated into the correct Chinese terms during the SPE process. This may have been due to differences in the types of user dictionary entries or the training data provided between the two languages; in any case, further investigation may reveal useful information for the effective use of SPE.

3 Sentence Level Evaluation

3.1 Evaluation setup

In addition to the aforementioned evaluation, we also conducted a pilot experiment evaluating the effect of SPE at the sentence level using three criteria: Fluency, Adequacy, and Post-Editing (PE) time. Based on the definition set by the Linguistics Data Consortium (LDC), fluency refers to the well-formedness based on the target language grammar, and adequacy refers to how much of the information and meaning of the original source text has been expressed in the target text (LDC 2005). These two criteria have been widely used (Papineni et al 2002, Turian et al 2003, Callison-Burch et al 2007, Doddington 2002, Owczarzak 2008, Boitet et al 2006), etc.

PE time here refers to the time needed to edit the output in order to raise the quality of the text so that it is appropriate for publishing purposes. PE time is important as PE is one of the major elements of human efforts in MT workflows and therefore reducing the PE time can have a significant impact on optimising the MT workflows.

Four evaluators for each language were recruited by Symantec. All four evaluators are native speakers of Chinese and Japanese respectively, and all are professional translators. Their experience in translation varies from three to twenty-two years, and the average is seven years. For each of the hundred segments, the English source text, RBMT output, and SPE processed text were presented to the evaluators, and they were asked to decide which target text is better than the other in terms of a) Fluency and b) Adequacy, and then decide which text would need less time to post-edit (Less-PE time). They were asked to put their answers in the table columns similar to Table 6 below. For each of the Fluency, Adequacy, and Less-PE columns, they were given three choices, 1 stands for the first output, 2 stands for the second output, while E stands for the equivalent quality for the two outputs. To avoid any bias on the part of the evaluators, the two target texts were presented in a mixed order, that is, for a random half of the hundred segments, RBMT output texts were presented as Output 1, and the SPE processed texts as Output 2, and the other way round for the rest of the segments. The four evaluators conducted the evaluation individually, and had no discussions or any form of information exchange during the evaluation.

Source_EN	Output 1	Output 2	Fluency	Adequacy	Less PE Time
Turns on or off the special meaning of metacharacters	オン/オフ回転メタ文字の特別な意味。	有効または無効にメタ文字の特別な意味します.	1 or 2 or E	1 or 2 or E	1 or 2 or E

Table 6. Human Evaluation Sample with Japanese output

We have used this simple "choice between three" method for the evaluation mainly due to time restrictions. Fluency and adequacy are normally evaluated in a scaled manner; for fluency, for instance, it is common that the evaluators are given five choices: 1: Incomprehensible, 2: Disfluent English, 3: Non-native English, 4: Good English, 5: Flawless English (Callison-Burch et al 2007, Boitet et al 2006), while we have given evaluators only a relative choice between MT, SPE, and Equal. The same is true for PE time. A number of attempts have been made to measure the post-editing effort using different methods (O'Brien 2007, Krings 2001), which have proved that measuring the post-editing effort is not a straightforward task. We are aware that the method we have employed here may put a restriction on supporting our findings.

3.2 Results and discussion

Table 7 shows the average results of the four evaluators for Chinese and Japanese. From the Chinese results, it can be seen that fluency and adequacy are regarded as having been improved during the SPE process in fewer than 40 cases on average, while PE time is thought to be shortened in nearly 50 cases. In contrast, for Japanese, fluency, adequacy, and PE time are considered to have been almost evenly improved in around 60 cases.

Language	Chinese			Japanese		
Criteria	Fluency	Adequacy	Less PE Time	Fluency	Adequacy	Less PE Time
MT	12.75	15.50	15.00	14.50	8.00	9.75
SPE	37.75	38.00	48.25	59.25	61.50	62.50
Equal	49.50	46.50	36.75	26.05	30.50	27.75
Total	100.00	100.00	100.00	100.00	100.00	100.00

Table 7. Average results for each criterion

After aggregating the results, we applied the Kappa coefficient equation (Carletta 1996) to the results to ensure inter-evaluator agreement, which showed another noticeable difference between the two languages. The following table shows the result of Kappa coefficients, which are widely used for determining the level of agreement among multiple evaluators (Callison-Burch et al 2007). According to the definition set by Landis & Koch (1977), 0.0 - 0.20 is regarded as having slight agreement, 0.21 - 0.40 fair agreement, 0.41 - 0.60 moderate agreement, 0.61 -

- 0.80 substantial agreement, and 0.81 - 1.00 almost perfect agreement. Based on this definition, Japanese inter-evaluator agreement is either at the higher level of moderate or the lower level of substantial agreement, while Chinese inter-evaluator agreement is all at the middle level of fair agreement. The high score for Japanese might be explained by the fact that the evaluation was conducted in a simple way by giving only three relative choices, but it does not explain the rather large difference in values between the languages. It must also be noted that low level of agreement might affect the generalisability of the findings.

Evaluation Criteria	Chinese	Japanese
Fluency	0.276	0.598
Adequacy	0.288	0.582
Less PE Time	0.284	0.624

Table 8. Kappa coefficient values for inter-evaluator agreement

3.2.1 Discussion on the Chinese results

The Chinese evaluators vary on their opinions; only 9 segments out of 100 gain unanimous judgements among the four evaluators and 36 gain unanimous judgements among three of the four evaluators. For the rest of the segments, different evaluators share different opinions on which are better in adequacy and fluency as well as which require less PE time.

Overall, the effect of SPE is obvious if we are looking for an incremental improvement in the translation quality. Many more credits were assigned to the SPE output than to the MT output. This may correlate with the improvements in different categories that we analyzed in section 2. Those improvements help to improve the adequacy, fluency of the translation and reduce post-editing time. For example, among the 7 sentences in which the four evaluators agree that SPE output is better in terms of Fluency and Adequacy and needs less PE time, 6 of them have at least one positive content words/phrases alteration.

However, degradation in those changes might also have a negative effect on adequacy and fluency and hence need more post-editing time. For 15% of the sentences, the evaluators think that the original MT output is better in adequacy and fluency and need less post-editing time. For example, for the three sentences which receive unanimous agreement that MT is better in fluency and adequacy and need less PE time, the degradations within them are inappropriate content or function words/phrases addition and inappropriate function words/phrases deletion.

3.2.2 Discussion on the Japanese results

One of the most striking outcomes found in the Japanese results is that, on average, the evaluators have estimated that SPE output should require less PE time in over 60% of the cases. In fact, in 42 segments out of 100, all four evaluators unanimously concluded that fluency, adequacy, and PE time have all been improved during the SPE process. The evaluators' opinions have varied in other cases where SPE time is considered to have been shortened. Therefore, it is not easy to conclude whether fluency improvement or adequacy improvement is likely to result in shorter PE time. In any case, it might be fair to say that, in general, SPE had a considerably positive impact on improving fluency, adequacy, and PE time.

On the other hand, there are eight segments where at least three evaluators have agreed that MT output should require less PE time than SPE output. Having investigated the reasons for this by revisiting the analysis conducted in section 3, it was found that one or more content word alterations had caused degradation during the SPE process in six out of eight cases. The remaining two cases consisted of one case where function word degradation occurred and another case where Words/Phrases Reordering caused degradation. This might suggest that controlling the content word alteration may, to some extent, help prevent adverse effects of SPE.

4 Conclusion and Future Research

In this study, we conducted a detailed investigation into the Chinese and Japanese results of a prior experiment carried out by Systran and Symantec (Senellart & Roturier forthcoming). The first objective of the research was to find out what linguistic changes SPE can and cannot make, and what their consequences are. The second objective was to research the effect of SPE in terms of fluency, adequacy, and reducing the subsequent human PE effort.

One of the notable findings from the linguistic analysis was that the most frequent changes made during the SPE process for both Chinese and Japanese were content words/phrases alterations, function words/phrases alterations, function words/phrases deletions, and punctuation changes. While content word alterations have resulted both in improvements and degradations in both languages, function word alterations, function word deletions, and punctuation changes have mostly resulted in improvements in both languages. It is an interesting finding that Chinese and Japanese share the same categories of changes, although the exact types of changes made within the same category differ partly due to the language differences. It may also be worth pointing out that the changes made during the SPE process are largely limited to the word level, and changes in the sentence structure or reordering of the words or phrases in a long range seemed difficult to achieve with the current SPE system.

Sentence level evaluation was also conducted to shed light on the effect of SPE in terms of fluency, adequacy, and PE time reduction. One of the most important findings from this evaluation is that the evaluators, on average, think the text after the SPE process requires less time for post-editing in around 50% and 60% of the cases in Chinese and Japanese respectively. This may suggest the potential of SPE in reducing the human effort in MT workflows, which could result in productivity gains, although the results are not clear-cut considering the simple form of evaluation method that was applied. Another curious finding is that the results of the sentence level evaluation contradict the results from the evaluation of changes conducted in Section 2. While SPE has a greater positive impact on Chinese than Japanese in the evaluation in Section 2, the sentence level evaluation has contradicted this and a noticeably better effect has been observed on the Japanese text. The result may be different if fluency and adequacy had been evaluated on a scale rather than with the “choice between three” method, and if the post-editing had been precisely timed, rather than subjectively assessed.

There are several limitations to the current study. Firstly, the conditions for two languages are not identical. Using the same RBMT system for both languages does not necessarily mean that the MT output quality and error types for the two languages are the same. By the same token, human translation in two different languages used for training SPE could not be guaranteed to have the same level quality. Moreover, the training and test materials used in this experiment for Japanese and Chinese were not identical. Secondly, the resources were limited. A detailed investigation was only carried out on a hundred sample segments, and only one native speaker of Japanese and Chinese (the authors) respectively worked on the quantitative and qualitative evaluation of the changes made by SPE in each language. In addition, although four evaluators participated in the sentence level evaluation in each language, because we used “a choice between three” method, rather than the scaled evaluation metrics, the data obtained is impressionistic. The evaluation for post-editing time is estimation rather than a strictly measured duration of editing time.

Nevertheless, this work may have revealed a number of possibilities and limitations of current SPE from a linguistic point of view, especially on less investigated languages such as Chinese and Japanese. A similar but larger scale research project could be conducted in the future using larger corpora with identical source text as well as using more finely scaled evaluation metrics for fluency and adequacy, and actual timing of post-editing. Also, comparison of the SPE text with the human post-edited text using some metrics to measure the textual differences, such as Translation Edit Rate (TER) (Snover et al 2006), may provide us with an interesting opportunity for further investigation of the correlation between the changes made during the SPE and their effects on PE effort.

References

Allen, J. & Hogan, C. (2000) Toward the Development of a Postediting Module for Raw Machine Translation Output: a Controlled Language Perspective. In: *Proceedings of The Third International Workshop on Controlled Language Applications (CLAW 2000)*, Seattle, Washington, pp. 62-71.

Boitet, C., Bey, Y., Tomokio, M., Cao, W. & Blanchon, H. (2006) IWSLT-06: Experiments with commercial MT systems and lessons from subjective evaluations. In: *Proceedings of International Workshop on Spoken Language Translation: Evaluation Campaign on Spoken Language Translation [IWSLT 2006]*, Kyoto, Japan, pp. 8-15.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. (2007) (Meta-) Evaluation of Machine Translation. In: *Proceedings of The Second Workshop on Statistical Machine Translation* Association for Computational Linguistics, , pp. 136-158.

Carletta, J. (1996) Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics,* 22, 2, 249-254.

Doddington, G. (2002) Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. In: *Proceedings of The Second International Conference on Human Language Technology*, San Diego, CA, pp. 138-145.

Dugast, L., Senellart, J. & Koehn, P. (2007) Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. In: *Proceedings of The Second Workshop on Statistical Machine Translation*, Prague, pp. 220-223.

Elming, J. (2006) Transformation-based correction of rule-based MT. In: *Proceedings of EAMT-2006: 11th Annual Conference of the European Association for Machine Translation*, Oslo, Norway, pp. 219-226.

Isabelle, P., Goutte, C. & Simard, M. (2007) Domain adaptation of MT systems through automatic post-editing. In: *Proceedings of MT Summit XI*, Copenhagen, Denmark, pp. 255-261.

J. R. Landis, J. R. & Koch, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics,* 33, 159-174.

Knight, K. & Chander, I. (1994) Automated Postediting of Documents. In: *Proceedings of 12th National conference of the American Association for Artificial Intelligence (AAAI 1994)*, Seattle, Washington, USA.

Koehn, P. (2004) Statistical Significance Tests for Machine Translation Evaluation. In: *Proceedings of Conference on Empirical Methods in Natural G128 Language Processing (EMNLP)*, pp. 388-395.

Krings, H.P. (2001) *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes.* The Kent State University Press, Kent, Ohio.

LDC (2005) *Linguistic Data Annotation Specification: Assessment of fluency and adequacy in translations.* Report number: Revision 1.5.

O'Brien, S. (2007) An Empirical Investigation of Temporal and Technical Post-Editing Effort. *Translation and Interpreting Studies,* 2, 1, 83-136.

Owczarzak, K. (2008) *A novel dependency-based evaluation metric for machine translation*, PhD edn, DCU.

Papineni, K., Roukos, S., Ward, T. & Zhu, W. (2002) BLEU: a Method for Automatic Evaluation of Machine Translation. In: *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, pp. 311-318.

Senellart, J. & Roturier, J. (forthcoming) Automation of Post-Editing in Localization Workflows. Presented at LISA Forum Europe 2008.

Simard, M., Goutte, C. & Isabelle, P. (2007a) Statistical Phrase-based Post-editing. In: *Proceedings of NAACL-HLT-2007 Human Language Technology: the conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, pp. 508-515.

Simard, M., Ueffing, N., Isabelle, P. & Kuhn, R. (2007b) Rule-based translation with statistical phrase-based post-editing. In: *Proceedings of ACL 2007: The Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, pp. 203-206.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. (2006) A Study of Translation Edit Rate with Targeted Human Annotation. In: *Proceedings of 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, pp. 223-231.

Turian, J.P., Shen, L. & Melamed, I.D. (2003) Evaluation of Machine Translation and its Evaluation. In: *Proceedings of MT Summit IX*, New Orleans, USA, pp. 386-393.

Vilar, D., Xu, J., D'Haro, L.F. & Ney, H. (2006) Error Analysis of Statistical Machine Translation Output. In: *Proceedings of LREC-2006: Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy, pp. 697.

Yoshimi, T. (2001) Improvement of Translation Quality of Pronouns in an English-to-Japanese MT System. *自然言語処理,* 8, 3, 87-106.