An Evaluation of Statistical Post-editing Systems Applied to RBMT and SMT Systems

Hanna Béchara[1,2] *Raphaël Rubino*[1]
Yifan He[3] *Yanjun Ma*[4] *Josef van Genabith*[1,2]
(1) NCLT, School of Computing, DCU, Dublin, Ireland
(2) CNGL, School of Computing, DCU, Dublin, Ireland
(3) CIMS, NYU, New York, USA
(4) Baidu, Beijing, China
hbechara@computing.dcu.ie, raphael.rubino@computing.dcu.ie,
yhe@cs.nyu.edu, yma@baidu.com, josef@computing.dcu.ie

ABSTRACT

Statistical post-editing (SPE) of the output produced by rule-based MT (RBMT) systems has been reported to produce extraordinary BLEU (and other automatic evaluation) score improvements. SPE has also been applied to the output of statistical MT (SMT) systems, albeit with more mixed results. We present a statistical post-editing pipeline and evaluate the outputs using automatic and human evaluation techniques, comparing the two SPE pipeline systems (RBMT + SPE and SMT + SPE) with the pure RBMT and SMT system, in an SPE scenario that uses independently existing bitext data, rather than manually corrected first stage MT output, as its training data. Our results show that although automatic evaluation metrics favour the pure SMT system, human evaluators prefer the output provided by the statistically post-edited RBMT system.

KEYWORDS: Rule-Based Machine Translation, Statistical Machine Translation, Statistical Post-Editing.

Proceedings of COLING 2012: Technical Papers, pages 215–230,
COLING 2012, Mumbai, December 2012.

1 Introduction

Human evaluation is a core component of shared tasks such as WMT, and is often considered the gold standard in evaluation of translation systems. Automatic evaluation metrics, however, are much less costly, much more time efficient and enable automatic tuning of SMT system parameters (e.g. using MERT), which may require a number of iterations to converge.

Statistical post-editing of the output of RBMT and SMT systems is an active field of research and RBMT + SPE pipelines are by now a commercial reality. [1] Automatic post-editing of rule-based machine translation systems (Simard et al., 2007; Terumasa, 2007; Kuhn et al., 2010) has shown (in some cases) spectacular improvements in translation quality measured in terms of automatic evaluation scores. Furthermore, SPE has also been applied to the output of statistical MT (SMT) systems (Oflazer and El-Khalout, 2007; Potet et al., 2011; Béchara et al., 2011; Rubino et al., 2012), albeit with more mixed results. However, to date, despite considerable interest in the area, the comparison between SPE pipelines and pure SMT and RBMT systems is not fully researched.

Previous research can be categorised into roughly two classes: in one approach (Simard et al., 2007; Dugast et al., 2007; Potet et al., 2011), manually corrected (i.e. post-edited) MT output is used as the target side for training the SPE system (i.e. a "mono-lingual" SMT system trained on the output of the first stage MT system as source and the manually corrected first stage MT output as target, and then applied to the output of the first stage MT system on unseen source side input data), while the other approach (Oflazer and El-Khalout, 2007; Béchara et al., 2011; Rubino et al., 2012) simply uses available bi-text training data (such as translation memories (TMs) in industrial applications or more generic SMT training data) and trains the SPE system on the output of the first stage MT system and the target side of the bi-text training data.

In the first approach, the SPE system is effectively trained in such a way as to only fix mistakes committed by the first stage MT system: the difference between the output of the MT system (the source side of the SPE training data) and the target side of the SPE training data is in translation mistakes identified and fixed by human post-editors. By contrast, in the second approach, the SPE system is simply trained on the difference between the MT output and the target side of the training data (which may not necessarily constitute a mistake, but just be an instance of a paraphrase or another valid translation alternative).

One of the central research questions addressed in this paper focuses on the second approach: do the (often spectacular) differences in automatic evaluation scores between RBMT and RBMT + SPE pipelines correspond to to "real" improvements in translation quality, as determined by human evaluators, or are they largely due to SPE moving RBMT output closer to a reference string used in automatic evaluation and rewarded by the automatic metrics?

Note that it is reasonable to assume that in the first approach improvements in automatic evaluation scores do indeed correspond to "real" improvements in translation quality, as determined by human evaluators, as in this case the SPE system is solely trained on genuine translation mistakes or (more generally) translation shortcomings.

At the same time, the second approach arguably addresses an important commercial reality in that in many industrial applications it is often the case that specialised TMs exist that have been developed (over many years) to support translation automation and that can now be used

[1] http://www.systran.co.uk/translation- products/server/systran-enterprise-server

as training data for SMT systems for translating input that yields only low fuzzy matches in the TMs. By contrast, the first approach involves a long term commitment to correct the first stage MT output to collect sizeable training data sufficient to train a successful SPE component, which is not always possible given deadlines and the variety of data that needs to be translated in commercial applications.

A second, related, research question addressed by the paper is to determine how RBMT + SPE pipelines compare with a pure SMT system (trained on the same data) or corresponding SMT + SPE systems (again trained on the same data), and how are these rated by human evaluators? Again we focus on this research question in the context of the second approach to SPE based on independently available bi-text training data such as TMs.

As automatic evaluation metrics are inherently biased by the reference translation, it is difficult to tell if these SPE scores correspond to an actual improvement in translation quality. In this study we use human evaluation to answer that question, and to further discover whether the statistical machine translation system outperforms the SPE combination systems, or whether the bias of the automatic evaluation metric is getting in the way of choosing the better system.

The remainder of this paper is organised as follows. In Section 2, we review related research and how it ties into our experiments. In Section 3 we detail the machine translation systems used in both the first and post-editing stages of our experiments, along with the automatic evaluation results of these systems and their combinations. Section 4 presents the human evaluation results, and Section 5 provides an error analysis that helps to answer some of the questions posed in this paper.

2 Related Work

Statistical post-editing has been applied to different types of MT systems to varying degrees of success. The main idea behind SPE for MT is to capture the mistakes made by the MT system and to automatically correct them. (Allen and Hogan, 2000) conducted early studies on the subject (without actually building a SPE system) by using a parallel corpus composed of three tiers: the source text, its automatic translation and the manually post-edited (i.e. corrected) automatic translation. This study inspired the original work on SPE by (Simard et al., 2007), who used the Portage System (a PBSMT system) to automatically post-edit the output of an RBMT system, using the raw RBMT output and the manually post-edited (i.e. corrected) output as "source" and "target" side, respectively, of the SPE training data.

SPE is generally mono-lingual, operating on the target side of the translation direction. In a sense, it can be viewed as a re-writing step on MT output a posteriori, usually unpacked as a supervised learning problem. In one approach, such as that of (Simard et al., 2007), SPE directly negotiates between specific errors in the RBMT (or generally first stage MT) output and the corresponding manual corrections (where mistakes in first stage MT output are corrected by human translators), in the other approach independent bitext data (such as TMs) are used to train SPE systems on the output of the first stage RBMT (or more generaly any MT) system on the sorce side of the bitext data and the corresponding target side of the bitext data. We present both approaches in the two subsections below:

2.1 SPE with Manually Post-Edited MT Output

Amongst the published work on SPE with manually post-edited MT output, several studies have been conducted by combining RBMT and PBSMT as the MT and the SPE systems, respectively.

(Simard et al., 2007) show that a commercial RBMT system combined with the PBSMT system Portage (Sadat et al., 2005) in an SPE pipeline achieve improved translation quality. On a translation task from French to English, using SPE shows an improvement of 13.7% BLEU (Papineni et al., 2002) (absolute) over the RBMT system alone. The authors also conduct experiments combining two PBSMT systems, both in the translation and the post-editing phase, and show that this approach leads to higher BLEU scores if the training corpora for the translation and the post-edition systems are different.

(Dugast et al., 2007) carry out a qualitative analysis at the linguistic level, conducted on the MT and the SPE outputs. They combine Systran with the PBSMT systems Moses (Koehn et al., 2007) and Portage. The output of the combined systems (Systran+SPE) shows significant improvements in terms of lexical choice. They also report gains in terms of BLEU scores up to 10 points absolute on a German to English translation task, compared to the RBMT system individually.

More recently, (Potet et al., 2011) combine a full PBSMT pipeline (SMT+SMT) for translation and post-editing from French to English. The first system translates the French text into English. The MT output is then manually post-edited and introduced into the pipeline following three approaches:

- as supplementary material to enrich the training corpus used to build the translation model,
- as the target side of the parallel corpus used to build the post-editing model,
- as a the target side of the development corpus used to optimize the translation model components weights.

This preliminary study shows a slight improvement over a standalone MT system, but further experiments on larger corpora are needed in order to obtain significant results.

2.2 SPE with Independent Bitext Data

Due to the time and expense involved in manually post-editing MT output, using independently available parallel training data (such as TMs) in SPE pipelines is often a less expensive (and arguably commercially more frequent) scenario. (Terumasa, 2007) combines RBMT with SPE to translate patent texts, and reports an improved score the NIST evaluation compared to that of RBMT alone.

In (Kuhn et al., 2010), the authors compare the two SPE approaches: the first using manually post-edited MT output and the second using the target side of the bilingual training data. They use Systran RBMT and Portage PBSMT systems, and combine them into a post-editing pipeline, with the RBMT system as first stage and the PBSMT system as the SPE system. The SPE system shows a gain of 10.2 BLEU points compared to the RBMT system alone, on a French-to-English translation task. However, the authors also show that a PBSMT system alone can reach results similar to those obtained by the post-editing pipeline.

The first mention of a pure SMT-based SPE pipeline (SMT + SPE) is likely to be in (Oflazer and El-Khalout, 2007), who in one of the experiments as part of their work on selective segmentation based models for English to Turkish translations, employ statistical post-editing (which they call model iteration). In the study conducted by (Béchara et al., 2011), two PBSMT systems are combined into a post-editing architecture. Two sets of experiments are presented. The first is a naive post-editing approach, were the output of the first SMT system is post-edited by the SPE

system without introducing additional information. The second is a source-context enriched SPE approach. Following the full PBSMT post-editing pipeline approach, (Rubino et al., 2012) use statistical post-editing to adapt out-of-domain machine translation systems to a specific domain and show that a generic MT system can be adapted through an automatic post-editing step.

In the first approach, SPE with manually post-edited MT output, the SPE directly addresses translation mistakes or inadequacies caused by the first stage MT system. By contrast, in the second approach, SPE with independently available bitext data, it is not guaranteed that a divergence between first stage MT output and the target side of the bilingual training data actually corresponds to a translation mistake by the first stage MT system. Our research focuses on this second scenario, in particular on whether (sometimes spectacular) SPE improvements in automatic evaluation scores correspond to actual improvements in translation quality verified by human evaluators. Our experiments follow the general design put forth by (Dugast et al., 2007) and (Kuhn et al., 2010), where the output of a Systran RBMT system on the source side of some bitext data as well as the target side of this bitext data is used to train a monolingual second-stage SPE system to post-edit the output of the RBMT system. The objective is to compare the output of the combined RBMT+SPE system with the RBMT system on its own, with an SMT system on its own trained on the bilingual training data, as well as with an SMT+SPE pipeline. We investigate whether the RBMT+SPE pipeline is actually better than Systran (or any of its other rivals: SMT and SMT+SPE), or just closer to the reference translation. In order to do this, we enlist the help of human evaluators.

3 Machine Translation Systems

In this section, we present the data and the translation systems used in our experiments. We also give details about the two SPE pipelines and the automatic evaluation metrics used to measure system performance.

3.1 Translation Memory Bitext Data

The data for our experiments are part of a French-English translation memory provided by Symantec. The data itself is in the domain of technical software user help information. We preprocessed the translation memory and removed all TMX markup and meta-information. We extracted 53,000 unique sentences from the translation memory, and from this data we randomly select 50,000 French-English sentence pairs as our training set. The sentences are between 1 and 98 words in length for English, and 1 and 100 words in length for French. The average sentence length in the training set is 13 words for English and 15 words for French, with a vocabulary size of 9,273 for the English side of the data, and 12,070 for French. Given the specific domain of the data, these are only 11 out-of-vocabulary (OOV) in the test set relative to the training data. The remaining sentences were split into a test set of 2000 sentences, and a development set of 1000 sentences. As we are working with a translation memory, all the sentences are unique. That is to say there are no repetitions in the data, and hence no overlap between the training, development, and test sets. Table 1 summarises the statistics of our data.

3.2 Rule-Based Translation

Despite the success of machine learning based and statistical approaches, rule-based machine translation systems still constitute a significant part of the current commercial MT landscape. RBMT systems work off built-in linguistic rules and bilingual dictionaries in order to construct

	French	English
TM Vocabulary Size	12,070	9,273
Training Vocabulary Size	11924	9159
Average Sentence Length	15	13
Range of Sentence Length	[1,100]	[1,98]

Table 1: Vocabulary and sentence length statistics for the French-English Translation Memory

translations for a given language pair. Wide-coverage systems rely on large-scale lexical and morphological, semantic, and syntactic information. Rule-based systems have been found to provide fluent and predictable quality translations.

As our rule-based machine translation system for the first stage MT in our experiments, we used the Systran Enterprise Server 6 production system, specifically customised with the use of 10K+ dictionary entries specific to the text type and domain of the Symantec translation memory data, as described in (Roturier, 2009).

3.3 Statistical Phrase-Based Machine Translation

Statistical machine translation builds statistical models based on the analysis of existing parallel corpora, both monolingual and bilingual. For our statistical machine translation system we used the PBSMT system Moses, 5-gram language models with Kneser-Ney smoothing trained with SRILM (Stolcke, 2002), the GIZA++ implementation of IBM word alignment model 4 (Och and Ney, 2003), with refinement and phrase-extraction heuristics as described in (Koehn et al., 2003). We used minimum error rate training (MERT) (Och, 2003) for tuning on the development set. During decoding, the stack size was limited to 500 hypotheses.

3.4 Statistical Post-editing

The first pipeline, which combines RBMT output with SPE (statistical post-editing) system, uses Systran to translate the entire source side of the TM-based training set, and the output together with the corresponding target side of the TM is then used as the training data for the SMT-based SPE system. The second-stage system therefore produces a monolingual translation based on the output produced by the first stage RBMT system and the target side of the TM training data (Figure 1).

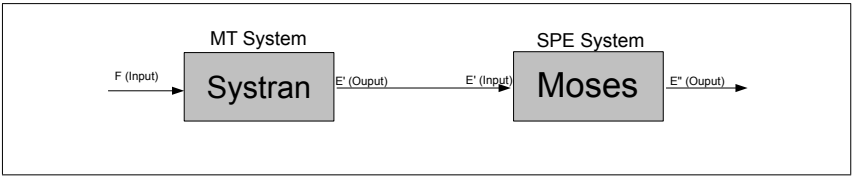

Figure 1: The RBMT+Moses pipeline, using the output of Systran as the input for the second stage SMT system (Moses)

The second pipeline uses SMT in both stages of the post-editing system. The first-stage PBSMT system is trained on the French to English parallel training data, and the output E' (MT output English) will be the source data for the second-stage (SPE) system. Once again the second-stage system produces a monolingual translation, this time using the output of the SMT system as its

input. The source side for the training data for the second-stage system is obtained by training another first-stage PBSMT system from French to English, using a 10-fold cross-validation approach on the French to English training set. This approach will ensure that we do not translate already seen data, and that the source side of the training set for the SPE system is as close in quality to the test set source as possible. Figure 2 illustrates the SMT+SPE pipeline.

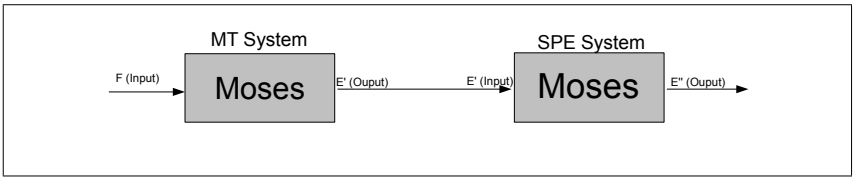

Figure 2: The Moses+Moses pipeline, using the output of Moses as the input for the second stage SMT system (Moses)

3.5 Automatic Evaluation Metrics

We used two metrics for automatic evaluation; BLEU (Bilingual Evaluation Understudy) and TER (Translation Edit Rate) (Papineni et al., 2002; Snover et al., 2006). Both of these metrics depend on a reference translation to estimate the quality of machine translated output. BLEU matches n-grams between the MT output and the reference translation, using n-gram precision with a brevity penalty as the score, as demonstrated in (1)

$$\text{BLEU}(n) = \prod_{1}^{n} PREC_i^{\frac{1}{n}} \cdot bp \tag{1}$$

where n is the order of n-gram, $PREC_i$ is the i-gram precision and bp is the brevity penalty. The brevity penalty is defined as (2):

$$bp = exp(max(\frac{len(Ref)}{len(Out)} - 1, 0)) \tag{2}$$

where $len(Ref)$ is the length of the reference and $len(Out)$ is the length of the output.
This n-gram matching scheme makes BLEU very sensitive to small changes in the output, and fails to capture linguistic variations, especially in the case where only one reference translation is being used. (Callison-Burch et al., 2008) show that that BLEU has a lower correlation with human judgement than metrics such as TER, which take into account linguistic resources and better matching strategies. Furthermore, BLEU is designed to evaluate MT output on a document level. For this reason, we have used S-BLEU (Sentence-Level BLEU) and TER to compare individual sentences.

TER is an Edit Distance-style evaluation metric that measures the amount of editing that a human post-editor would have to perform to change a system output so it matches the given reference translation. It calculates how many insertions, deletions, substitutions and sequence shifts are required to make the output identical to the reference. TER is defined in equation(3):

$$\text{TER} = \frac{\#INS + \#DEL + \#MOD + \#SHIFT}{len(Ref)} \tag{3}$$

	RBMT	SMT	RBMT+SPE	SMT+SPE
BLEU	23.26	65.43	64.63	65.14
TER	61.07	23.92	24.62	24.12

Table 2: BLEU and TER scores for the RBMT, SMT and the SPE systems

System	Ins	Del	Sub	Shift	TER
SMT	5.1	5.05	10.5	3.5	23.92
RBMT	17.04	4.39	30.24	9.3	61.07
SMT+SPE	5.47	4.95	10.1	3.56	24.61
RBMT+SPE	5.2	5.5	10.5	3.27	24.11

Table 3: Normalised number of translation errors for the RBMT, SMT, and SPE systems according to TER edit statistics

3.6 Automatic Evaluation Results

In order to evaluate the RBMT, SMT, RBMT+SPE and SMT+SPE approaches, we train and tune the SMT system on the French-to-English training and development sets, and decode the 2,000 sentence test set using this first-stage system. We also translate the complete training data using a 10-fold cross training approach (tuning on the same development set as above) to avoid translation of seen data to create the source side of the SPE system for the SMT+SPE pipeline. Furthermore, we translate the complete training data (as well as the test set data) using the Systran RBMT system, to create the source side training data for the SPE system in the RBMT+SPE pipeline. The target side of both SPE systems is provided by the target side of the training data from the TM. We evaluate automatically the post-edited translation of the test set using BLEU and TER. Despite the fact that the RBMT system was tuned to the TM data (section 3.2), Table 2 shows that SMT, RBMT+SPE and SMT+SPE outperform RBMT by more than 40 BLEU points absolute on our TM data set (with similar improvements for TER).[2] While the RBMT+SPE system improves over the RBMT output, it fails to improve over the pure SMT output, and performs similar to the SMT+SPE output in terms of the automatic evaluation scores.

Table 3.6 presents the number of average edits per sentence, based on the TER edit types. The errors are divided into four categories: insertion (Ins), substitution (Sub), deletion (Del) and shift. The numbers have been normalised using sentence length to make them comparable. The table suggests that applying SPE to the RBMT system achieves significant gains in the insertion and substitution categories, and to a lesser extent to the shift category. This reflects the fact that the SPE system can improve the pure RBMT translation in terms of better lexical choice and better reordering. Furthermore, the large number of substitutions and insertions in the RBMT system shows that the majority of the errors that account for the lower quality of the RBMT system are lexical. The number of deletions remains largely unaffected by the post-editing system, indicating that little information is actually lost during the second stage. Neither the RBMT+SPE nor the SMT+SPE systems achieve any significant gains over the pure SMT system.

4 Human Evaluation

In order to further investigate the quality of the RBMT+SPE output compared to the pure RBMT, pure SMT as well as the SMT+SPE output, we complement automatic evaluation metrics with a study involving human evaluators. Human evaluation can be an important source of

[2]Note again that there is no duplication between the test and training data extracted from the TM.

information, providing insight as to whether or not (and why) the SMT system actually performs better or worse than the RBMT+SPE pipeline, how these compare with the SMT+SPE output, and whether, and if so to what extent, the spectacular differences in automatic scores between the RBMT and the SMT and SPE pipeline systems actually correspond to human judgements.

4.1 The Evaluation Task

Evaluation was carried out by ten different translators of varied backgrounds. While none of them are professional translators, all of them have experience with machine translation or localisation. Six of these evaluators are native French speakers, and the others have a good grasp of French, evidenced by school and professional certificates. All of the evaluators speak English fluently.

The evaluators were asked to evaluate a pair of sentences from two of the four MT systems: pure SMT, pure RBMT and the two SPE pipelines RBMT+SPE and SMT+SPE. The evaluators were shown a source sentence (in French) and asked which of the two MT outputs (presented in random order) is a better translation, or if they are of equal quality. In order avoid biasing the evaluator, we did not provide a reference translation. The task spanned 200 sentences, and was available to be completed online. The subjects were paid for their time and were given a week to submit the task, which did not have to be completed in one sitting. Figure 3 shows a screenshot of the user interface with an example sentence included in the task. The evaluators generally rated the task as difficult, especially as the domain was highly technical and the sentences often fragmented and containing a large number of symbols and abbreviations.

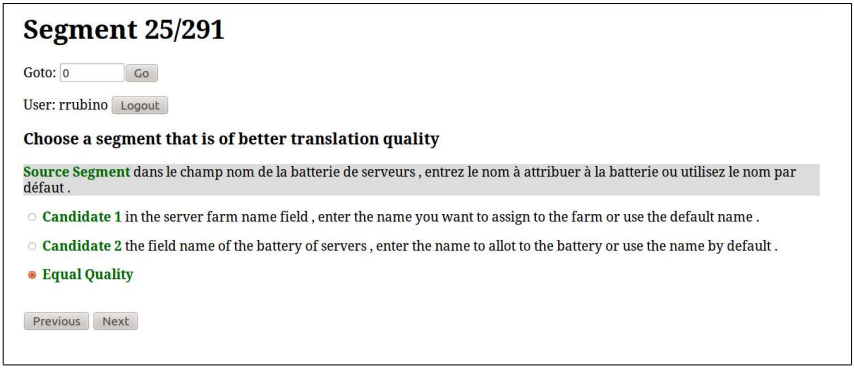

Figure 3: A screen-shot of the manual evaluation task

4.2 Annotator Agreement

Since a reasonable degree of agreement must exist to support the validity of our human evaluation experiment, we calculated pair-wise inter-annotator agreement between all of the different evaluators. For this agreement, we used Cohen's κ measure. κ is a more robust measure compared to simple percent agreement calculation, as it takes into account the

agreement occurring by chance. κ is defined by the formula in 4.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{4}$$

Pr(a) is the proportion of times two annotators are observed to agree, and *Pr(e)* is the expected proportion of times the two annotators are expected to agree by chance. Agreement occurs when two annotators compare the same systems and agree on their rankings. In our case there are three possible choices; either one system is better than the other, or it is worse, or there is a tie. κ ranges between 0 and 1, with 1 indicating a higher rate of agreement, and 0 indicating low or no agreement.

According to (Landis and Koch, 1977), a moderate agreement falls between 0.4 and 0.6. A substantial agreement falls between 0.6 and 0.8, and 0.2 to 0.4 indicate a fair agreement, while anything below that is considered slight. Full results for all ten evaluators ($\kappa = 0.42$) are on the border between moderate to fair. As two of our evaluators scored an average agreement under 0.4, we discarded their results as weak. Without the outliers, our average agreement for the evaluators is 0.47. This amounts to low moderate agreement.

4.3 Human Evaluation Results

We tallied up the number of times that each system was chosen as "best system" based on the 200 sentences that were evaluated. The human annotators' results were normalised and divided by the number of annotators (in our case 8, as we discarded the outliers). In order to evaluate the document on a sentence-level, we used S-BLEU instead of BLEU. S-BLEU will still positively score segments that do not have highern-gram (n=4 in our setting) matching, unless there is no unigram match.

We compared the S-BLEU and TER scores for the sentences of each of the outputs, and tallied up the number of times each system was given the best score by either S-BLEU or TER. We then compared the results from the human evaluators with the S-BLEU and TER results. We found that while the automatic metrics seem to favour a pure SMT system, the human annotators were leaning more towards the combined rule based and post-editing system. Furthermore, while automatic evaluation metrics chose Systran as the best system less than 7% of the time, human evaluators chose it as the best system more than twice as often as S-BLEU or TER did. The results are further detailed in Table 4.

In order to assess the difficulty to compare different systems, we recorded the time each evaluator spent evaluating a translation pair. We assume that spending more time on an evaluation indicates that it is more difficult to select the best translation amongst the two alternatives. We report averaged results in Table 5. The results show that comparing the two stand-alone MT systems (RBMT and SMT) is hard because on average more than 20 seconds are spend to select the best translation. This is most likely due to the profound differences in terms of syntax and vocabulary between the SMT and the RBMT outputs. By contrast, when comparing SMT versus SMT+SPE, the time spent drops by nearly 10 seconds (on average). This is most likely due to the fact that SMT and SMT+SPE outputs are very similar, requiring less time to scan and judge. A similar trend can be observed comparing SMT with RBMT+SPE, where again outputs are more similar than between SMT and RBMT on their own. Finally, choosing between RBMT and RBMT+SPE requires the least amount of time. This is consistent with the observation that (according to the human evaluation) the quality difference between

	Human Evaluation	S-BLEU	TER
	SMT vs RBMT		
SMT	97	162	161
RBMT	52	16	9
Tie	51	26	30
	SMT vs RBMT+SPE		
SMT	28	125	124
RBMT+SPE	40	50	46
Tie	132	25	30
	SMT vs RBMT+SPE		
RBMT	40	16	11
RBMT+SPE	99	162	162
Tie	61	22	26
	SMT+SPE vs RBMT		
SMT+SPE	107	167	161
RBMT	46	11	9
Tie	47	25	30
	SMT+SPE vs RBMT+SPE		
SMT+SPE	27	46	46
RBMT+SPE	47	49	41
Tie	126	105	113

Table 4: Number of sentences chosen as "best" by each of the evaluations

RBMT and RBMT+SPE is the most pronounced, and therefore more "obvious" than in the other cases.

	SMT vs RBMT	SMT vs SMT+SPE	SMT vs RBMT+SPE	RBMT vs RBMT+SPE
Average time	23.4	14.12	12.61	9.52

Table 5: Average time spent (in seconds) by human evaluators on each system comparison.

5 Error Analysis

In order to obtain a better understanding of the translation quality gains between the RBMT system and the RBMT+SPE system, and to gain insight into why there are discrepancies between the manual and automatic evaluation results, we performed an additional manual sentence-level error analysis in a bid to reveal the advantages and disadvantages of the SPE pipelines compared with the RBMT and SMT systems.

Table 6 shows the detailed number of error types by system, based on the error typography provided by (Vilar et al., 2006). Our error analysis confirms what the TER edit statistics in Table 3.6 suggest, that most of the errors that account for the considerably lower quality of the RBMT system are lexical, both in terms of simple lexical choice and the repercussions of this on the phrasal level. Even though the RBMT system was tuned to the domain of the TM via domain specific lexical resources (Section 3.2), most of the errors appear to be due to the RBMT system's inability to pick the right term for the technical domain data set. However, compared to SMT and SMT+SPE, both the RBMT and RBMT+SMT system seem to produce a significantly lower number of grammatical errors, according to our evaluators. This is mostly obvious in the determiner and preposition categories, where combined, the SMT system produces three times as many errors as the RBMT system. Our results also show that while the SPE considerably

	RBMT	SMT	RBMT+SPE	SMT+SPE
Not Found Words	1.5	5	0	5
Simple Terms	34.5	10.5	6	9.5
Phrases	20.5	2.5	2	3
Meaning	20.5	2.5	2	3
Determiners	1	4.5	2	2.5
Prepositions	3	8.5	2.5	6
Tense	1.5	2.5	2.5	3
Number	0	1	1	1
Other Grammar	2.5	6.5	3.5	5.5
Punctuation	1	3.5	3.5	3.5
Word Order	7	4	4	4.5

Table 6: Normalised number and types of errors found in manual evaluation results

changes the error typography when applied to RBMT, reducing the overall number of errors, it has has a much smaller effect when applied to the SMT system. SMT+SPE fails to improve on a lexical choice where SMT has failed, and only marginally improved grammatical errors. Example 1 shows a very common RBMT lexical choice error. Errors such as these are almost always corrected in the statistical post-editing (SPE) phase.

Example 1

Source	options de planification de modification d'a pour **un travail de sauvegarde**
RBMT	options of planning of modification of has for **a work of backup**
RBMT+SPE	schedule options to change for **a backup job**
SMT	scheduling options has changed for **a backup job**
SMT+SPE	scheduling options has changed for **a backup job**
Reference	to change schedule options for for **a backup job**

Example 2 shows a similar case where the RBMT+SPE pipleline is superior when it comes to picking the right phrases within the correct domain. Due to the highly technical nature of the translation memory, the intended meaning is often lost if the wrong lexical choices are made.

Example 2

Source	pour installer **une version d évaluation**
RBMT	to install **a version of rating**
RBMT+SPE	to install **a trial version**
SMT	to install **a trial version**
SMT+SPE	to install **a trial version**
Reference	to install **an evaluation version**

On the other hand, RBMT often performs better when it comes to general grammar, especially in terms of prepositions and, to a lesser extent, determiners. This carries over to the RBMT+SPE system, which leads to a better grammatical quality than the pure SMT system (or the SMT+SPE pipeline, for that matter). Example 3 shows a common case where the preposition is missing from the Moses translation, but is inserted correctly in both the RBMT and RBMT+SPE translations.

Example 3

Source	pour ajouter le **le nom de compte** de connexion
RBMT	to add the name **of account** of login
RBMT+SPE	to add the name **of logon account**
SMT	to add the name **logon account**
SMT+SPE	to add the name **logon account**
Reference	to add the **logon account** name

Another interesting aspect concerns out-of-vocabulary (OOV) words. RBMT seems to be better at finding words than SMT (this is probably a reflection of the fact that the RBMT system used in our experiments was a production system tuned with a domain-specific 10k+ dictionary to the TM-based data-set), and even though these are not always perfectly correct words, they are sometimes fixed in post-editing, as seen in Example 4. As a result, RBMT and RBMT+SPE produce few if any out of vocabulary items given the output.

Example 4

Source	**enregistrera** l'image .iso idr amorçable ou non amorçable
RBMT	will **record** the or not bootable image .iso idr bootable
RBMT+SPE	will **save** the idr bootable or non-bootable .iso image
SMT	**enregistrera** the idr bootable or non-bootable .iso image
SMT+SPE	**enregistrera** the idr bootable or non-bootable .iso image
Reference	will **save** the bootable or non-bootable idr .iso image .

The results also show that in a few cases SMT+SPE can produce some grammatical improvements over the pure SMT system as well. Example 5 is one such case, where SPE applied to SMT corrected a grammatical error (the missing preposition **for**).

Example 5

Source	le nombre de secondes **pendant lesquelles le processus de restauration** ...
RBMT	the number of seconds **during which the process of restoration** ...
RBMT+SPE	the number of seconds **for the restore process** ...
SMT	the number of seconds **the restore process** ...
SMT+SPE	the number of seconds **for the restore process** ...
Reference	the number of seconds **for the restore process** ...

Conclusion and Perspectives

Previous research on automatic post-editing has shown spectacular improvements in automatic MT evaluation scores of RBMT + SPE pipelines over RBMT systems. In this paper we set out to answer two open questions for an SPE scenario that uses independently existing bitext training data, rather than specifically and manually corrected first stage MT output, as its training data:

- Is statistical post-editing (SPE) improving the output of the first-stage RBMT system, or is it just pushing the output closer to the reference translation?
- Does the RBMT+SPE pipeline improve the quality of the output over that of the pure SMT system as well as that of an SMT+SPE system trained on the same data set?

In order to answer these questions we used human evaluation from annotators who were unbiased by the reference translation. Our human evaluators agreed with the automatic evaluation metrics that the RBMT + SPE system does indeed perform better than the RBMT system on its own. Additionally, while they did not find the improvement as pronounced as the automatic evaluation metrics indicate, they consistently rated the RBMT + SPE system higher than the RBMT system by a factor of two.

While automatic evaluation metrics indicate that SPE does not improve RBMT systems over pure SMT system, a manual evaluation shows that human evaluators prefer the RBMT + SPE output over the pure SMT output. We conclude that this discrepancy is a result of S-BLEU and TER being biased towards the SMT system. Error analysis shows that SPE makes better lexical and phrasal choices, which leads to superior translation quality. We also observed that the human evaluators spend more time to select the best translation when the MT systems are very different (SMT and RBMT for instance), which reflects the underlying difficulty of the evaluation task.

We would like to use what we learnt about the nature of the errors and the strengths and weaknesses of the post-editing system to automatically predict which sentences can be corrected using SPE. Many sentences do not benefit from the post-editing phase, and a subset of sentences even degrade after post-editing. Using machine learning techniques, we plan to classify the output of the RBMT system, based on a variety of features, into two categories: those that benefit from SPE, and those that do not. Furthermore, we plan to train a classifier to choose better sentences of the two systems, RBMT+SPE and pure SMT, in order to reach the upper bound given by the two systems together. Furthermore, results obtained from the error analysis should enhance feature selection methods.

We will research ways to further refine statistical post-editing techniques for both RBMT and SMT systems. In previous work, we developed a contextualised SPE system that attempts to preserve the original context of the source material with a novel method of context modelling (Béchara et al., 2011). We intend to take this work further and refine our use of context information. We will also experiment with different system combinations: in addition to RBMT and PBSMT systems, we will utilise a hierarchical phrase based SMT system (Chiang, 2005).

Acknowledgments

This work is supported by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. We would like to thank Symantec, in particular Dr. Johann Roturier and Dr. Fred Hollowood, for providing us with the data and with access to the Symantec Systran Machine Translation Production System. We would like to thank Dr. Stephen Doherty, from the School of Applied Language and Intercultural Studies at Dublin City University, for providing us with access to masters in translation students to act as our evaluators. We would also like to thank the reviewers for their many insightful comments, suggestions, and corrections.

References

Allen, J. and Hogan, C. (2000). Toward the Development of a Post editing Module for Raw Machine Translation Output: A Controlled Language Perspective. In *Proceedings of the Workshop on Controlled Language Applications (CLAW)*, pages 62–71.

Béchara, H., Ma, Y., and van Genabith, J. (2011). Statistical Post-Editing for a Statistical MT System. In *Proceedings of the MT Summit XIII*, pages 308–315.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further Meta-Evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT)*, pages 70–106.

Chiang, D. (2005). A Hierarchical Phrase-based Model for Statistical Machine Translation. In *ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*.

Dugast, L., Senellart, J., and Koehn, P. (2007). Statistical Post-Editing of SYSTRAN's Rule-Based Translation System. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT)*, pages 220–223.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 177–180.

Koehn, P., Och, F., and Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 48–54.

Kuhn, R., Isabelle, P., Goutte, C., Senellart, J., Simard, M., and Ueffing, N. (2010). Automatic Post-Editing. *Multilingual*, 21(1):43–46.

Landis, J. and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Och, F. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 160–167.

Och, F. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. In *Proceedings of the Association for Computer Linguistics (ACL)*, pages 29(1):19–51.

Oflazer, K. and El-Khalout, I. D. (2007). Exploring Different Representational Units in English-to-Turkish Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT)*, pages 25–32.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 311–318.

Potet, M., Esperança-Rodier, E., Blanchon, H., and Besacier, L. (2011). Preliminary Experiments on Using Users' Post-Editions to Enhance a SMT System. In *Proceedings of the European Association for Machine Translation (EAMT)*, pages 161–168.

Roturier, J. (2009). Deploying Novel MT Technology to Raise the Bar for Quality: A Review of Key Advantages and Challenges. In *Proceedings of the MT Summit XII*.

Rubino, R., Huet, S., Lefèvre, F., and Lenarés, G. (2012). Statistical Post-Editing of Machine Translation for Domain Adaptation. In *Proceedings of the European Association for Machine Translation (EAMT)*, pages 221–228.

Sadat, F., Johnson, J., Agbago, A., Foster, G., Kuhn, R., Martin, J., and Tikuisis, A. (2005). PORTAGE: A Phrase-Based Machine Translation System. In *Proceedings of the Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*.

Simard, M., Goutte, C., and Isabelle, P. (2007). Statistical Phrase-based Post-editing. In *Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 508–515.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and J., M. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA)*, pages 223–231.

Stolcke, A. (2002). SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 901–904.

Terumasa, E. (2007). Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation. In *Proceedings of the MT Summit XI Workshop on Patent Translation*, volume 11, pages 13–18.

Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 697–702.