Towards a Better Understanding of Statistical Post-Edition Usefulness

Marion Potet[1], *Laurent Besacier*[1], *Hervé Blanchon*[2], *Marwen Azouzi*[1]

[1] UJF-Grenoble1, [2] UPMF-Grenoble2
LIG UMR 5217, Grenoble, F-38041, France
FirstName.LastName@imag.fr

Abstract

We describe several experiments to better understand the usefulness of statistical post-edition (SPE) to improve phrase-based statistical MT (PBMT) systems raw outputs. Whatever the size of the training corpus, we show that SPE systems trained on general domain data offers no breakthrough to our baseline general domain PBMT system. However, using manually post-edited system outputs to train the SPE led to a slight improvement in the translations quality compared with the use of professional reference translations. We also show that SPE is far more effective for domain adaptation, mainly because it recovers a lot of specific terms unknown to our general PBMT system. Finally, we compare two domain adaptation techniques, post-editing a general domain PBMT system *vs* building a new domain-adapted PBMT system with two different techniques, and show that the latter outperforms the first one. Yet, when the PBMT is a "black box", SPE trained with post-edited system outputs remains an interesting option for domain adaptation.

1. Introduction and Related Work

The post-edition task consists of editing the textual output produced by an error-prone process (Machine Translation, Optical Character Recognition, Speech Recognition, etc.) in order to improve it. In documents diffusion workflows where Machine Translation (MT) is one of the components, manual post-edition has been used for years. The MT system produces raw translations (or translation hypotheses) which are manually post-edited by professional translators or trained post-editors who correct the translation errors.

Many studies have shown the benefits of using MT combined with manual post-edition in a diffusion workflow. The work presented in [1] showed that even if post-editing raw MT output does not lead to any improvement in terms of productivity, it helps to produce significantly better translations compared to direct manual translations from the source text, regardless of the language direction, the text difficulty or the translator's experience. Autodesk recently draw opposite conclusion of an experiment to test whether using MT would improve translators' productivity or not. Indeed, the results[1] showed that post-editing MT output leads to a significant increase in productivity when compared with translations done from scratch, whatever the language pair, the experience and preference (post-editing or translating from scratch) of the translator, or the sentence length.

Improving the quality of the output in terms of fluency and adequacy has always been a major goal of MT developers, and in the manual post-edition setting, *"the better the MT output, the easier and faster post-edition will be"*. In the early 90's, K. Knight and I. Chander [2] proposed automated post-edition (APE) in order to help with article selection when translating from Japanese to English. Later, J. Allen and C. Hogan [3] proposed the development of an automated rule-based post-edition module able to capture and correct *"the frequent and repeated errors produced by Rule-Based Machine Translation (RBMT) systems*. Then, J. Elming [4] was the first to propose and evaluate an APE module. In his settings, J. Elming carried out domain-specialized translations of chemistry patents, cascading a RBMT system called *Patrans*, used to produce raw translations, with a "transformation-based" APE trained on 12 000 manually post-edited translations, to correct the raw output. There was a significant improvement in translation quality with the use of a "transformation-based" APE. The increasing amount of raw MT translation (hypotheses) aligned with their manually post-edited good translations gave rise to the idea of automatic statistical post-edition. A statistical post-edition (SPE) system is developed as a monolingual statistical MT system using the original hypotheses as the source language and the human post-editions as the target language.

In 2007, M. Simard & al. [5] were the first to propose the use of a phrase-based statistical machine translation (PBMT) system for SPE purpose. In this framework, the PBMT aims to learn "correction rules" between initial MT hypotheses (PBMT source language) and their corrected version (PBMT target language). Such an approach makes SPE easy to learn and tune with new training data. In their work, they successfully showed the efficiency of using an SPE system (built with the PBMT *Portage*) to improve the output of a commercial RBMT system. The experiments were done in a specific domain (a job offer Web site[2]) and the SPE system was trained using 35,000 manually post-edited sentences. Encouraged by these results, post-editing the outputs

[1] http://translate.autodesk.com/productivity.html

[2] http://www.jobbank.gc.ca

of the PBMT system *Portage* was also tried but in this setting no improvements were observed. In the same way, the following studies described in [6], [7] and [8] have shown that a RBMT system that was automatically post-edited by a PBMT system performed significantly better than each of the individual systems on their own.

Quite a lot of studies have focused on pipeline architectures where SPE systems are successfully applied to RBMT systems outputs to improve translation quality. However, only few studies ([9, 8, 10]), have investigated the efficiency of SPE systems applied after PBMT systems.

The goal of our study is to provide a better understanding of SPE usefulness when pipelined to PBMT systems. We first describe our baseline experimental settings (Section 2) and then we try to answer the following questions: is there a difference between a real and a simulated corpus for SPE training (Section 3)? Is SPE useful in improving a generic PBMT system and what explains the effectiveness of SPE on specialized domain (Section 4)? And, finally, is SPE really the simplest and most efficient and effective way for domain-adaptation purposes (Section 5)?

2. Experimental setting

2.1. Baseline PBMT

Our baseline MT system (described in more detail in [11]) translates news stories (general domain) from French into English. It is a state-of-the-art phrase-based machine translation (PBMT) system presented at the international Workshop of Machine Translation (WMT[3]) evaluation campaign in july 2010.

The system was built using free open source toolkits: we used standard Moses [12] system set-up, a 3-gram language model trained with SriLM [13] and Kneser-Ney smoothing, the GIZA++ implementation of IBM word alignment model 4 [14] and the phrase extraction heuristics described in [12]. The system has been trained on two parallel corpora, containing in total 1,638,440 aligned sentences: the fourth version of the Europarl corpus (data derived from transcriptions of European parliament proceedings) and news corpora (data extracted from various Websites). Both corpora were provided in the framework of WMT 2010.

The PBMT decoding model is a log-linear combination of fourteen weighted feature functions extracted from the monolingual and bilingual training data: six distortion models; lexicon word-based and phrase-based translation models for both directions; a target language model; and word, phrase and distortion penalty models.

2.2. Post-edited corpus

Our parallel post-edited corpus is a set of 10,881 French/English sentences taken from several news corpora (WMT evaluation campaigns from 2006 to 2010). Each sentence has been translated with our baseline PBMT system and the translation hypotheses have been manually post-edited by human annotators who were given the French source sentence and its English translation hypothesis and had to verify the translation quality and correct it if needed.

Post-editions were collected using a crowdsourcing Web platform (Amazon Mechanical Turk - MTurk). The ethical, social and economic aspects implied by such tools are subject to intense debates [15], so we defined and applied the following "good conduct" guidelines: data collected for the contributors should be used for non-profit organization and available for free to the community; contributors should be informed about the context of the task (Who are we? What are we doing? And why?); contributor should be paid a decent amount (with a reasonable hourly rate); and contributors should be filtered by country of residence according to the task, to avoid those who consider MTurk as their major source of income (we only authorized American, Canadian, and French residents to participate in our study).

Contributors were required to have an understanding of the French language and be fluent in English. Clear instructions and controlled review allowed us to deal with untrained human post-editors (native of the target language or not). A complete analysis of the collected data indicated high quality corrections with more than 94 % of the crowdsourced post-editions which are at least of professional quality. Some examples of translation hypothesis corrections collected during the post-edition campaign are given in Table 1. The post-editions corpus collection and data analysis are more detailed in [16].

The collected corpus was divided into three subsets: 8,681 sentences for the SPE training set, 1,000 sentences for the SPE development set, and 1,200 sentences for the SPE test set. Thus, all the following SPE experiment results are evaluated on the 1,200 sentences long test corpus.

For each French source sentence, we have our English baseline PBMT translation hypothesis and two different reference translations: the baseline post-edited output and an independent professional translation provided with the parallel corpus.

2.3. Baseline SPE system

As in many of the previous experiments reported here, we have considered automatic post-edition as a translation task performed by a PBMT system where the source corpus consists of the raw MT outputs and the target corpus consists of the post-edited version of these raw translations.

Our SPE system was developed using the same architecture and the same tools we used for our baseline system (Moses, SriLM and GIZA++). We trained the SPE models on the training set of the post-edited corpus (8,681 sentences) and adjusted the model's features weights with the Minimum Error Rate Training (MERT) process [17] on the development set of the post-edited corpus (1,000 sentences).

The language model was trained on a general domain cor-

[3] http://www.statmt.org/wmt10/

Source Sentence	PBMT translation	PBMT + human corrections
• La police anti-émeutes les ont aussitôt encerclés et sont intervenus sans ménagement, jetant plusieurs d'entre eux à terre.	• The anti-riot police were immediately surrounded and spoke blunty, several of them on land.	• ~~The~~ **A**nti-riot police**men** ~~were~~ immediately surrounded **them** and ~~spoke blunty~~ **stepped in ruthlessly**, **throwing** several of them ~~on land~~ **to the ground**.
• Forte mobilisation à Copenhague et à travers le mode, pour le climat.	• Strong involvement in Copenhague and in the world climate.	• Strong ~~involvement~~ **mobilization** in Copenhague and ~~in~~ **across** the world **for the** climate.
•Il y a des rivières qui s'assèchent en Afrique, des cours d'eau où l'on peut marcher comme on ne l'avait jamais fait avant.	• There are rivers are drying up in Africa, rivers where you can walk as it had never done before.	• There are rivers ~~are~~ drying up in Africa, ~~rivers~~ **watercourses** where ~~you~~ **one** can walk as it had never done before.

Table 1: Examples of PBMT hypothesis post-editions

pus of 48,653,884 english sentences (about 2 billion words).

The result is a phrase table where English baseline SMT output segments are aligned with their corresponding human post-edition. As a statistical translation model, the SPE system takes as input a raw MT output and produces a new translation hypothesis using its models.

2.4. Evaluation metrics

Translation output quality has been evaluated using the Translation Error Rate (TER) [18] and the BLEU score [19]. The TER score reflects the number of edit operations (insertions, deletions, words substitutions and blocks shifts) needed to transform a hypothesis translation into the reference translation, while the BLEU score is the geometric mean of n-gram precision. Lower TER and higher BLEU scores suggest better translation quality. To ensure that differences between scores are real, we estimated the statistical significance of test results in terms of BLEU score, according to the bootstrap resampling method described in [20].

3. Real *vs* Simulated post-edited corpus for SPE training

3.1. Previous work

In order to build SPE systems, manually post-edited MT hypotheses are usually used as target translations instead of translations produced by professional translators. When pre-existing human translations are used, we will speak of "simulated PE" in contrast to "real PE" when target translations are manually post-edited MT hypotheses. It is important to notice that the "real PE" setting corresponds to the workflows implemented in real-life situations (when users feedback is re-used to improve a given system) and "simulated PE" setup will allow access to much more training data (use of pre-translated parallel corpus).

Several works [21, 10, 22, 9] have attempted to show that SPE can be successfully trained on pre-existing human translations rather than on system-specific post-edited translations. Both simulated (MT system hypotheses aligned with their human translations version) and real post-edited (MT system hypotheses aligned with their manually post-edited versions) training corpora are used in [23]. Each setting ("real" SPE and "simulated" SPE) shows good results, but performances are not really comparable because neither the RBMT system baseline nor the SPE training corpus (in terms of size and domain) are the same in the two cases.

To our knowledge, there is no work that compares both approaches (real *vs* simulated PE) on the same source language data (post-edited MT hypotheses *vs* professional translations) to train an SPE. Considering the same source language data, we tried to find out if a simulated PE corpus is as effective as a real PE corpus to train an SPE system. This is what we will try to find out in the following experiment.

3.2. Experiment

In order to build two comparable SPE using real *vs* simulated target corpus, we used in both cases the same training corpus on the source side (the one described in 2.2) and, for one system we used the PBMT post-edited hypotheses ("real" setting) on the target side and for the other system, we used the translations provided with the parallel corpus ("simulated" setting) as the target side. Both SPE were applied on the same PBMT system outputs and we estimated the translation quality of each SPE on the test corpus (1,200 sentences) using the same distinction as we did for the training corpus: we used the test set post-edited MT outputs, for the "real" setting, and the professional translations for the "simulated" setting.

System	Simulated PE corpus	Real PE corpus
PBMT	55.3 (*26.5*)	22.8 (*62.1*)
PBMT + SPE	57.5 (*25.0*)	23.4 (*61.3*)

Table 2: Performance — TER (*BLEU*) scores — according to the use of the simulated *vs* the real post-edited corpus to train the SPE

3.3. Results

As presented in Table 2, raw PBMT output obtains a TER score of 22.8 when compared with human post-editions and 55.3 when compared with independent reference translations. A TER score of 22.8 means that slightly over 22.8% of the words needed to be changed to produce the "correct" (or reference) translation.

We expected that real post-edited corpus would lead to better results than the simulated one because of the closeness between MT raw translation hypotheses and translation post-editions. Applying the "real" SPE on PBMT outputs led to a slight increase of the TER (from 22.8 for PBMT outputs to 23.4 after statistical post-editing) and decrease of BLEU score (from 62.1 for PBMT outputs to 61.3 after statistical post-editing). However, these differences in scores do not reach a significant level (according to [20]).

So, the SPE system trained on real post-edited corpus does not significantly degrade translation results, whereas there is a significant deterioration when post-editing with the SPE trained on simulated post-edited corpus (after statistical post-editing, translation quality loses relatively 4.0% of TER score and 6.0% of BLEU score).

According to our experiment settings (i.e. a medium size corpus and general domain data), we noticed that statistical post-edition of our PBMT system brings no improvement whatever the data (real *vs* simulated) used for SPE training.

3.4. Is more data always better?

To complete our previous result, we studied the impact of training corpus size on the SPE performance. Given the moderate size of our available human post-edited corpora (10,881 sentences), we considered simulated SPE to carry out larger-scale experiments.

We used the French/English United Nation parallel corpus which consists of the texts of resolutions made by the UN General Assembly, translated by professionals. In the SMT translation community, this corpus is widely used as a general and large training corpus[4].

We considered the 8,681 sentence-sized (10k) news corpora (see part 2.2) and split the UN corpus to set up a 50,000 sentence-sized (50k), 100,000 sentence-sized (100k), 500,000 sentence-sized (500k), 1,000,000 sentence-sized (1M) and 2,000,000 sentence-sized (2M) corpora (each included the 10k news corpus). We then trained SPE systems on those 6 corpora. Note that the only thing that differentiates the systems is the training corpus size. The LM used in the different sized experiments is the same as the one used by the baseline SPE system in Section 2.3.

We evaluated the different SPE systems on the test set and report the performances, in terms of TER and BLEU scores, on Figure 1 (systems are ranked according to their training corpus size). The results show no significant gains,

[4]The corpus is available at http://www.statmt.org/wmt12/translation-task.html

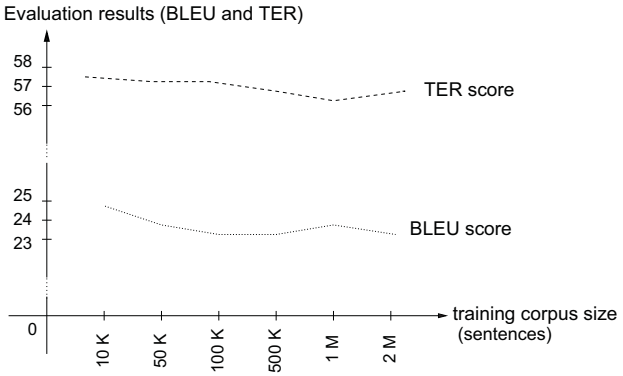

Figure 1: Performance — TER (*BLEU*) scores — of simulated SPE systems according to training corpora size (in sentences)

neither for the TER score nor for the BLEU score, while the corpus size increase. In other words, in a general French/English context translation, additional training data do not improve the SPE result of our PBMT system.

4. General domain *vs* Domain-specific application for SPE

4.1. Previous work

As SPE has shown its effectiveness in significantly improving RBMT results, further works have focused on its application in domain adaptation. Thus, P. Isabelle & al. [6] and M. Simard & al. [7] showed that an SPE trained on domain-specific data could be used to adapt a general RBMT system to a new specialized domain.

D. De Ilarraza & al. [8] noticed that if applying an SPE system after a RBMT system is efficient enough to adapt the RBMT system to a new domain, applying an SPE system after a PBMT system, for the same task, does not lead to any improvement. In their works, A. Lagarda & al. [10] and H. Becahara & al. [9] reached the same conclusion when they applied a baseline domain-specific SPE on generic PBMT system outputs. The work presented in [9], meanwhile, proposed some SPE customizations, by adding the source context into the post-edition to improve PBMT domain-adaptation.

Even if these studies confirm SPE efficiency when applied after RBMT for domain adaptation purpose, they do not show positive results when an SPE system is applied after a PBMT system. As shown before in our study, general-domain SPE brings no improvement when applied after a generic PBMT system. If the SPE system could not correct the PBMT system, can an SPE system be used to adapt the same baseline system to a new domain? To answer this question, we set up an experiment to test the potential of a generic SPE approach compared to a domain-specific one.

4.2. Experiments

Given the nature of our available corpora, the following experiments use only a simulated post-edited corpus for SPE training. We used the post-edited corpus described in 2.2 with the independent professional reference translations and a domain-specific corpus on water sciences.

The domain-specific and general corpora used for our experimentations are described in Table 3. They are very comparable in terms of size and only differ from each other by their domain specificity. As the general domain corpus, the domain-specific corpus has been split into a training set ($\approx$ 9,000 sentences), a development set (1,000 sentences) and a test set (1,200 sentences). A new SPE system has been built using the domain-specific data (the previous one presented used general domain data).

4.3. Results

As seen in Table 4, the general domain baseline PBMT achieves a TER score of 55.3 on the general domain and a score of 46.7 on the specific domain, meaning that these latter data are easier to translate than those of the general domain. Although the general domain SPE brings no gain on general data, the specific-domain SPE significantly improves the baseline PBMT outputs on the specialized data: the TER score subsequently drops from 46.7 to 39.2 (-19.2%) and the BLEU score follows the same trend, increasing from 33.3 to 40.1 (+20.6%).

The first line of Table 5 indicates that the domain-specific SPE is not only better (as seen in Table 4) but it modifies more sentences (91%) as compared to the general domain SPE (which modifies 75% of sentences). The second line shows the proportion of baseline PBMT translations improved through statistical post-edition: the specific-domain SPE improves 58% of the PBMT outputs while only 11% for the general domain SPE. Some examples of domain-specific translations before and after post-editions are presented in Table 6.

System	Specific domain	General domain
PBMT	46.7 (*33.3*)	55.3 (*26.5*)
PBMT+SPE	39.2 (*40.1*)	57.5 (*25.0*)

Table 4: Systems' performances — TER (*BLEU*) scores — according to the domain

4.4. Real domain adaptation or vocabulary correction?

The main follow up questions raised by these new experiments are: Why does SPE work on the domain-specific inputs and fail on general ones? Is SPE doomed to domain-adaptation? In [21], SPE modifications in the raw MT output have been manually categorized and results conclude that SPE makes significant improvements in terms of lexical choice, but no improvement in word reordering or grammaticality.

[5]http://www.statmt.org/wmt10

Post-edit rate	Specific domain	General domain
Post-edited sentences	91 %	75 %
SPE-improved PBMT outputs	58 %	11 %

Table 5: Rate of post-edited sentences according to the domain

Is SPE successful in domain-adaptation task only thanks to lexical correction? We decided to analyze how SPE handles out-of-vocabulary (OOV) words. So, we compared OOV words before and after general and domain-specific SPE. We did this experiment on two sets of 2,200 sentences (concatenated development and test sets for both domain-specific and general domain settings).

The results, shown in Table 7, point out an equivalent proportion of OOV words in both sets (2.8% for the domain-specific corpus and 2.7% for the general one) but with a type-token OOV word ratio[6] of 61 %, the domain-specific data contain less lexical variation than the general one. The application of SPE corrected 56% of the PBMT outputs OOV words for the domain-specific data and 7% for the general data.

OOV words statistics	Specific domain	General domain
Outputs with OOV words	40 %	43 %
Rate of OOV words	2.8 %	2.7 %
Type-token OOV words ratio	61 %	72 %
OOV words corrected by SPE	56 %	7 %
OOV common nouns corrected by SPE	**42%**	**1%**

Table 7: OOVs statistics according to the domain

Nature of corrected OOV words	Specific domain	General domain
Proper nouns	16.8 %	**46.8 %**
Foreign language words	2.3 %	34.7 %
Source mistake	1.5 %	2.4 %
Numbers	3.3 %	5.6 %
Common nouns	**75.6 %**	9.7 %

Table 8: Nature of corrected OOVs according to the domain

In order to better understand these results, we analyzed the nature of OOV words for both data sets. The results

[6]The type-token ratio is a measure of text vocabulary variability. The higher is the ratio, the larger is the lexical variability.

Corpus	Specific	General
Domain	Water Sciences	News
Nature	EOLSS encyclopaedia	Various websites
Vocabulary size	14 015 words	21 982 words
Sentence length	$\approx$ 22 words	$\approx$ 28 words
Source	Corpus translated by SECTra_w project [24]	Corpus provided by WMT international workshop[5]

Table 3: General *vs* specific corpus comparison

Source sentence	PBMT translation	PBMT + SPE result
• Unité africaine de recherche sur les questions de l'eau	• African **unit of research on issues of water**	• African **water issues research unit**
• Réduction de la salinité des eaux souterraines dans les zones agricoles	• Reducing **the salt content of groundwater** in agricultural areas	• Reducing **groundwater salinity** in agricultural areas
• L'offre est en grande partie déterminée par la productivité dans les zones irriguées et pluviales[...]	• **The offer** is largely determined by productivity in the irrigated **areas and *pluviales***[...]	• **Supply** is largely determined by productivity in the irrigated **and rain-fed areas**[...]

Table 6: Examples of specific-domain translations

are presented in Table 8. We noticed that the baseline PBMT OOV words are mostly common nouns (75.6%) for the domain-specific data, whereas they are mostly proper nouns and foreign language words (81.5%) for the general data. In a translation task, the latter just have to be copied out (this is what the baseline PBMT usually does with OOV words) whereas common nouns have to be correctly translated. The figure to retain is that SPE corrects 42% of OOV common nouns on the domain-specific data and only 1% on the general data.

OOV correction analysis also showed that the SPE learned to correct very domain-specific words that frequently appear in the data (for example: ions, évaporite, électrolytes, etc.). Our experiment results indicate that, when applied to domain specific data, SPE corrects a lot of OOV common nouns. This can explain the overall translation quality improvement. To sum up: SPE does not safely and effectively correct a general PBMT system output but it does some good work for domain adaptation thanks to its ability to restore domain-specific vocabulary. The follow up question remains: Is another simple domain adaptation method capable of outperforming SPE?

5. Domain-specific SPE *vs* other domain-adaptation methods

As SPE seems to be an efficient domain-adaptation method, we propose to compare this approach to other usual domain-adaptation methods. For these experiments, we used the general domain data and the PBMT system described in Section 2 and the domain-specific data described in Section 4.

5.1. Corpus-based domain-adaptation experiments

Our corpus-based domain-adaptation method consists simply of appending the domain-specific corpus to the general domain training corpus and then build the PBMT system as usual. The success of this straightforward method depends on the homogeneity of both corpora, i.e. the way they complete one another (in terms of OOV coverage, for example) and basically on the relative size of both corpora. As seen in Table 9 line (2), we get a significant improvement in terms of BLEU and TER (+37.0% and -25%) despite the fact that the general domain data greatly outnumbers the domain-specific one (which represents only 0.5% of the total training corpus). However, we reached better improvement by giving greater weight to the domain-specific training data by appending it several times to the corpus used for training (results line (3), (4) and (5)). The system achieved its best performance in terms of BLEU and TER (+48.2% and -45.0%) with domain-specific data weighing 35.5% of the total corpus size (line (4)).

5.2. Model-based domain-adaptation experiments

Corpus-based domain-adaptation methods led to a huge increase in the training time. Instead of simply concatenating all of the available training data, we have experimented with two methods using multiple phrase tables (PT) and language models (LM).

On one hand, we built separate phrase tables and language models for each data sets (domain-specific LM and PT, general domain LM and PT) and then we used all of them in the log-linear model. This model-based adaptation method is referred to "*domain-specific PT-LM*$_1$", line (6) in Table 9.

On the other hand, we tried to interpolate specific and general language models before using it in the log-linear

Baseline PBMT	...with domain-specific SPE	...with domain-specific PT-LM$_2$
• There is some maximum quantity of water vapor for each of the value of the air temperatures. • This is in connection with the effects of noise. • A reduction in consumption of animal products will very probably a positive effect on consumption of water to agriculture	• There is **some maximum** amount of water vapor for **each of the** value of the air temperature. • This is in **connection with the effects of acoustic**. • A **shift** in consumption of **animal products** will **most likely** positive effect on water consumption **to** agriculture	• There is **a certain** amount of water vapor **maximum possible** for **every** value of the air temperature. • This is in **relation to the acoustic effects**. • A **reduction** in **the** consumption of **products of animal origin** will **very probably** a positive effect on water consumption **of** agriculture

Table 10: Examples of translations according to the domain-adaptation method

Systems	TER(*BLEU*)
Generic PBMT	*46.7 (33.3)*
(1) domain-specific SPE	39.2 (*40.1*)
—— Corpus-based adaptation ——	
(2) 1×domain-specific corpus (=0.5%)	35.2 (*45.5*)
(3) 10×domain-specific corpus (=5.2%)	33.1 (*48.5*)
(4) 10^2×domain-specific corpus (=35.5%)	32.3 (*49.2*)
(5) 10^3×domain-specific corpus (=84.5%)	32.6 (*48.9*)
—— Model-based adaptation ——	
(6) domain-specific PT-LM$_1$	33.0 (*47.9*)
(7) domain-specific PT-LM$_2$	32.2 (*49.2*)

Table 9: Performance — TER (*BLEU*) scores — on a specialized domain corpus according to domain adaptation method

model. The LMs interpolation weights were estimated using an EM algorithm[7] and then, the two LMs were merged (using SriLM tool [13]) into a single model. We observed a slight improvement in terms of BLEU and TER (referred as "*domain-specific PT-LM$_2$*", line (7) in Table 9).

According to the experiment results, the systems produced with the corpus-based and the model-based domain-adaptation methods (TER from 32.2 to 35.2) significantly outperform the SPE method (TER of 39.2). Figure 10 shows some examples of specific-domain translation hypotheses using the domain-specific SPE system and the *domain-specific PT-LM$_2$* system.

6. Conclusion

The aim of this study was to better understand the usefulness of statistical post-edition to improve PBMT systems outputs. In order to do so, we tried to answer the following questions: Is simulated SPE really comparable to real SPE? Can an SPE system be applied to PBMT system outputs in order to improve them? Can an SPE system be used to adapt a general domain "black-box" MT system towards a particular domain? For domain-adaptation, is SPE more efficient than building a new domain-adapted PBMT sytem?

First, we noticed that an SPE system trained on moderate-size and general domain data ($\approx$ 9,000 sentences) brings no gain to a baseline general domain PBMT system in terms of TER or BLEU. In such a setting, using manually post-edited outputs ("real setting") instead of independent professional reference translations ("simulated setting") leads to a slight improvement of the translation quality. We also observed that increasing the amount of the training data is not sufficient to significantly improve the SPE system performances. So, whatever the available corpora, it seems difficult to improve/correct, general domain PBMT outputs with statistical post-editing.

However, according to our experiments, an SPE system seems more effective when trained on domain-specific data and can be successfully used to adapt a general PBMT system to a new specialized domain. Comparing our general domain and domain-specific SPE systems, we pointed out that better results are achieved with the latter one. This is mainly due to the fact that in-domain unknown common nouns of the general-domain PBMT system are recovered by the domain-specific SPE system.

In our last experiment we decided to compare SPE-based domain-adaptation with another adaptation approach which consist of training specialized phrase-tables and language models and interpolate them with the baseline general models. For this latter experiment, each methods shared the same baseline PBMT system and the same data sets. Results show that the PT-LM domain-adaptation method significantly outperforms the domain-specific SPE.

It is however important to note that in the case of model-based adaptation, a brand new PBMT system is built. There might be practical situations where it is impossible to build a new PBMT system (the one used is a "black box"), or it may be useful to keep a general PBMT system and a record of several SPE systems each adapted to a different domain.

7. References

[1] I. Garcia, "Translating by post-editing: is it the way forward?" *Journal of Machine Translation*, vol. 25, no. 3,

[7] http://sourceforge.net/apps/mediawiki/irstlm/index.php?title=LM_interpolation

pp. 217–237, 2011.

[2] K. Knight and I. Chander, "Automated Postediting of documents," in *Artificial Intelligence Conf.*, Seattle, USA, 1994, pp. 779–784.

[3] J. Allen and C. Hogan, "Toward the development of a post-editing module for Machine Translation raw output: a controlled language perspective," in *International Controlled Language Applications workshop*, Washington DC, USA, 2000, pp. 62–71.

[4] J. Elming, "Transformation-based corrections of rule-based MT," in *European Association on Machine Translation Conf.*, Oslo, Norway, 2006, pp. 219–226.

[5] M. Simard, C. Goutte, and P. Isabelle, "Statistical phrase-based post-editing," in *North American Chapter of the Association for Computational Linguistics and Human Language Technologies conf.*, Los Angeles, USA, 2007, pp. 507–515.

[6] P. Isabelle, C. Goutte, and M. Simard, "Domain adaptation of MT systems through automatic post-editing," in *North American Chapter of the Association for Computational Linguistics*, Stroudsburg, USA, 2007, pp. 217–220.

[7] M. Simard, N. Ueffing, P. Isabelle, and R. Kuhn, "Rule-based translation with statistical phrase-based post-editing," in *Statistical Machine Translation workshop*, Prague, Czech Republic, 2007, pp. 203–206.

[8] A. Diaz de Ilarraza, G. Labaka, and K. Sarasol, "Statistical post-editing: a valuable method in domain adaptation of RBMT systems for less-resourced languages," in *Mixing Approaches to Machine Translation*, Donostia-San Sebastian, Spain, 2008, pp. 35–40.

[9] H. Béchara, Y. Ma, and J. Van Genabith, "Statistical post-editing for a statistical MT system," in *MT SUMMIT XIII*, Xiamen, China, 2011, pp. 308–315.

[10] A. Lagarda, S. Casacuberta, and E. Diaz-de Liano, "Statistical post-editing of a Rule-based machine Translation System," in *North American Chapter of the Association for Computational Linguistics conf.*, Boulder, Colorado, USA, 2009, pp. 217–220.

[11] M. Potet, l. Besacier, and H. Blanchon, "The LIG machine translation system for WMT 2010," in *Statistical Machine Translation workshop*, Uppsala, Sweden, 2010, pp. 11–17.

[12] H. Hieu, A. Birch, C. Callison-burch, R. Zens, R. Aachen, A. Constantin, M. Federico, N. Bertoldi, C. Dyer, B. Cowan, W. Shen, C. Moran, and O. Bojar, "Moses: Open source toolkit for statistical machine translation," in *Association for Computational Linguistics*, Prague, Czech Republic, 2007, pp. 177–180.

[13] A. Stolcke, "SRILM: An Extensible Language Modeling Toolkit," in *Spoken Language Processing conf.*, Denver, USA, 2002, pp. 901–904.

[14] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Journal of Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[15] K. Fort, G. Adda, and K. B. Cohen, "Amazon Mechanical Turk: Gold Mine or Coal Mine?" *Journal of Computational Linguistics*, vol. 37, pp. 413–420, June 2011.

[16] M. Potet, E. Esperan¸ca Rodier, L. Besacier, and H. Blanchon, "Collection of a Large Database of French-English SMT Output Corrections," in *Language Resources and Evaluation Conf.*, Istanbul, Turkey, 2012, pp. 23–25.

[17] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Association for Computational Linguistics*, Sapporo, Japan, 2003, pp. 71–79.

[18] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Association for Machine Translation in the Americas conf.*, Cambridge, USA, 2006, pp. 223–231.

[19] K. Papineni, S. Roukos, T. Ward, and Z. Wei-Jing, "BLEU : A Method for Automatic Evaluation of Machine Translation," in *Association for Computational Linguistics*, Philadelphia, PA, USA, 2002, pp. 311–318.

[20] P. Koehn, "Statistical Significance Tests for Machine Translation Evaluation," in *Empirical Methods in Natural Language Processing conf.*, Barcelona, Spain, 2004.

[21] L. Dugast, J. Senellard, and P. Koehn, "Statistical post-editing on Systran's rule-based translation system," in *Statistical Machine Translation workshop*, Prague, Czech Republic, 2007, pp. 220–223.

[22] L. Dugast, J. Senellart, and P. Koehn, "Statistical post-editing and dictionary extraction: systran/edinburg submissions for WMT2009," in *Statistical Machine Translation workshop*, Athens, Greece, 2009, pp. 110–114.

[23] R. Kuhn, P. Isabelle, C. Goutte, J. Senellart, M. Simard, and N. Ueffing, "Recent advances in automatic post-editing," *Journal of Multilingual computing and technology*, vol. 21, no. 1, pp. 43–46, 2010.

[24] C.-P. Huynh, C. Boitet, and H. Blanchon, "SECTra_w.1: an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora," in *Language Resources and Evaluation Conf.*, Marrakech, Morocco, 2008, pp. 28–30.