Statistical Post-Editing of Machine Translation for Domain Adaptation

Raphaël Rubino **Stéphane Huet** **Fabrice Lefèvre** **Georges Linarès**

LIA-CERI
Université d'Avignon et des Pays de Vaucluse
Avignon, France
{firstname.lastname}@univ-avignon.fr

Abstract

This paper presents a statistical approach to adapt out-of-domain machine translation systems to the medical domain through an unsupervised post-editing step. A statistical post-editing model is built on statistical machine translation (SMT) outputs aligned with their translation references. Evaluations carried out to translate medical texts from French to English show that an out-of-domain machine translation system can be adapted *a posteriori* to a specific domain. Two SMT systems are studied: a state-of-the-art phrase-based implementation and an online publicly available system. Our experiments also indicate that selecting sentences for post-editing leads to significant improvements of translation quality and that more gains are still possible with respect to an oracle measure.

1 Introduction

Phrase-Based Machine Translation (PBMT) is a popular approach to Statistical Machine Translation (SMT) that leads to accurate translation results (Zens et al., 2002; Marcu and Wong, 2002; Koehn et al., 2003). The statistical models used in PBMT are based on the probabilities of bidirectional alignment of phrases between two sentences in the translation relation. The linguistic resources used to estimate such probabilities are parallel corpora and the main resulting statistical model is a translation table. Therefore, parallel corpora are the cornerstone for high quality translation. However, such resources are expensive to construct.

This lack of parallel data still remains an issue in PBMT. This phenomenon is accentuated by the diversity of texts to translate, in terms of origin and domain. As explained in (Sager et al., 1980), most of human activities involve a specific language or a *subject language*. A specific domain can be characterized by particular terminology or syntactic and discourse structures. As building domain specific translation systems for each domain is unreasonable, we assume that domain adaptation of out-of-domain translation systems can be one of the solutions to address the diversity of specific domains.

Although current machine translation systems can lead to impressive accuracy, translated texts require sometimes human post-processing to be usable. However, editing *a posteriori* can be costly depending on the amount of corrections required by machine translation outputs. Therefore, the automation of post-editing is an important task which can lead to higher quality machine translation without requiring human intervention.

In this paper, we propose a statistical post-editing (SPE) approach to adapt SMT systems to specific domains. We focus on translating texts in the medical domain from French to English. Several SMT systems are studied and we propose different methods to include the in-domain data into the translation process. We evaluate how translation quality can be improved with a post-editing step based on a phrase-based alignment approach. Two sets of experiments are presented in this paper: one applying SPE consistently on all the sentences and one resorting to SPE only on selected sentences.

The remainder of this paper is organized as follows. Section 2 presents the phrase-based post-editing approach. In Section 3, we propose an ex-

perimental setup and give details about the data, the language models, the translation and the post-editing systems used in our experiments. Section 4 evaluates each SMT system on a domain specific translation task, then Section 5 analyses the effect of a standard post-editing system on translated texts. Section 6 presents our approach to select sentences for post-editing. Finally, Section 7 concludes this paper.

2 Phrase-Based Statistical Post-Editing

2.1 SPE Principles

The post-editing of a machine translation output consists of the generation of a text T'' from a translation hypothesis T' of a source text S. When a PBMT system is built on bilingual parallel data, a phrase-based SPE system requires monolingual parallel texts. Recent approaches on SPE are based on three-part parallel corpora composed of a source language text, its translation by an MT system and this output manually post-edited (Knight and Chander, 1994; Allen and Hogan, 2000). If SPE can correct mistakes made by machine translation systems, it can also be used to adapt machine translation outputs to specific domains.

2.2 SPE for Adaptation

The research presented in this paper addresses the issue of adapting an out-of-domain machine translation system using a small in-domain bilingual parallel corpus. We study various uses of out and in-domain data to build Language Models (LMs) and Translation Models (TMs) inside the source-to-target language PBMT. Then, we evaluate the post-editing model using out and in-domain data to build LMs and in-domain data only for the SPE model. We also describe a new method to select sentences using classifiers built with the BLEU criterion (Papineni et al., 2002).

Figure 1 illustrates the general architecture of our experimental setup, described in the next section. The source language part of the in-domain parallel corpus is first translated into the target language by an SMT system. Then, the generated translation hypotheses are aligned with their translation references in order to form a monolingual parallel corpus and to build a SPE model. When a test corpus is translated and has to be post-edited, we propose two different approaches. The first one is a *naive* application of SPE which post-edits all the sentences of the test corpus. The second one is based on a classification approach that aims to avoid a degradation of translation quality at the sentence level. For this last approach, we build a sentence classification model to predict whether or not the sentences from the test set can be improved with SPE.

2.3 Related Work

In (Simard et al., 2007a), the authors propose to post-edit translations from a Rule-Based Machine Translation (RBMT) system using the PBMT system PORTAGE (Sadat et al., 2005). A qualitative study of phrase-based SPE is presented by (Dugast et al., 2007; Dugast et al., 2009), where the Systran system outputs are post-edited with PORTAGE and MOSES. The authors report gains up to 10% absolute of BLEU.

In (Isabelle et al., 2007; Simard et al., 2007b), it is shown that a generic, or out-of-domain, RBMT system can be adapted to a specific domain through phrase-based SPE. Domain specific data are introduced at the post-editing level, which globally improves the translation quality. Besides, de Ilarraza et al. (2008) propose the same architecture, phrase-based SPE following a RBMT system, and introduce a small amount of in-domain data to train the SPE model, as well as morphological information in both systems.

More recently, Béchara et al. (2011) design a full PBMT pipeline that includes a translation step and a post-editing step. The authors report a significant improvement of 2 BLEU points for a French to English translation task, using a novel context-aware approach. This method takes into account the source sentences during the post-editing process through a word-to-word alignment between the source words and the target words generated by the translation system. This latter work is, to the best of our knowledge, the first attempt to combine two PBMT systems, one for translating from the source to the target language, and another one for post-editing the first system output.

This kind of PBMT pipeline had already been suggested by previous authors (Isabelle et al., 2007; Oflazer and El-Kahlout, 2007). Let us note that their work is not targeting to improve the outputs of an out-of-domain SMT system with adaptation data as in our approach. Another recent approach related to our work was presented in (Suzuki, 2011) to select sentences for post-editing. The authors present an architecture

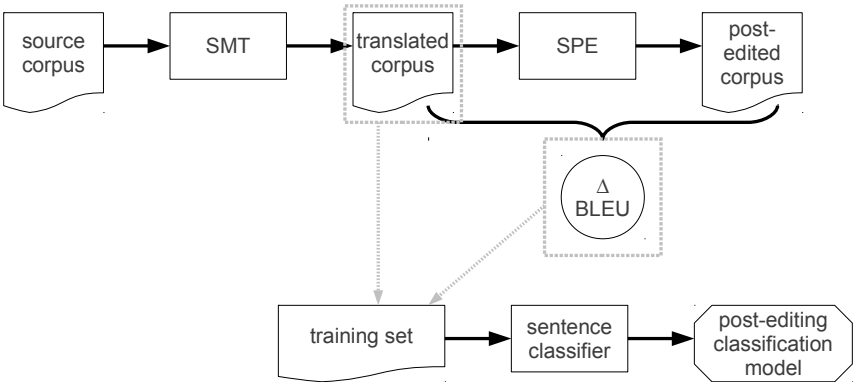

Figure 1: Training of a SVM classifier using a translated corpus where each sentence is associated with its $\Delta BLEU$ class.

composed of a phrase-based SPE system and a sentence-level automatic quality estimator based on Partial Least Squares.

3 Experimental Setup

In brief, the general idea of the work presented in this paper is to increase the quality of in-domain translations, generated by an out-of-domain SMT system, through a post-editing step. In order to thoroughly evaluate our approach, two SMT systems are considered to translate from the source language to the target language: the MOSES PBMT implementation (Koehn et al., 2007) and the GOOGLE TRANSLATE online system[1]. The post-editing step is then performed using MOSES in both cases. The latter case (the online system) will help to justify our approach showing that a powerful yet fixed MT system can be profitably combined with a system trained on a small set of in-domain data. The approach is evaluated at two levels: first, we evaluate the accuracy of each translation system on a domain specific translation task. Second, we focus on the use of SPE systems to process each translation system output.

Section 3.1 introduces the out and in-domain data used in our experiments. These data can be combined in different ways inside LMs and TMs; the resulting translation systems are described in Section 3.2. Then, Section 3.3 provides information about our SPE models.

3.1 Resources

Out-of-domain data are presented in Table 1. The bilingual parallel corpora are the sixth version of the Europarl corpus (Koehn, 2005) and the United Nations corpus (Rafalovitch and Dale, 2009). The monolingual corpora are composed of the target language part of the sixth version of the News Commentary corpus taken from the *Project Syndicate* website[2], and the Shuffled News Crawl corpus. All these corpora were made available for the 2011 Workshop on Machine Translation (WMT11)[3]. The bilingual data are used to build translation models, whereas the monolingual data are employed to train language models.

Corpus	Sentences	Words
Bilingual Training Data		
Europarl v6	1.8 M	50 M
United Nations	12 M	300 M
EMEA (Medical)	160 k	4 M
Monolingual Training Data		
News Commentary v6	181 k	4 M
Shuffled News from 2007 to 2011	25 M	515 M

Table 1: Number of sentences and words for the out and the in-domain data used in our experiments.

The in-domain domain data used in our experiments are taken from the EMEA corpus (Tiedemann, 2009), made out of PDF documents from the European Medicines Agency[4]. The source documents are associated with three biomedical categories: general medical documents and public evaluation reports about human or veterinary treatments. This corpus is particularly interesting because it contains medical terminology and specific linguistic structures. Since the EMEA corpus contains lots of repeated expressions (on med-

[1] http://translate.google.com/

[2] http://www.project-syndicate.org/

[3] http://www.statmt.org/wmt11/

[4] http://www.emea.europa.eu/

ical prescriptions for instance), we removed duplicates. Furthermore, short sentences of one word and long sentences exceeding 80 words were discarded. The resulting corpus is split separately for each category into three parts, which globally leads to three corpora: a 156k-sentence training set, a 2k-sentence development set and a 2k-sentence test set.

3.2 Initial SMT Systems

The online translation tool, noted com in the remainder of this paper, cannot be modified. It provides us with translation hypotheses which can be scored and post-edited in order to evaluate our approach. The MOSES PBMT implementation can be used to train a translation model from parallel corpora. Several PBMT systems are built, based on the bilingual and monolingual data used.

Three different 5-gram Kneser-Ney LMs are trained on the resources, using the SRILM toolkit (Stolcke, 2002). A first one (LM_g) is built on the monolingual out-of-domain data while a second one (LM_m) is built on the target language part of the medical (in-domain) corpus. These two models are combined through a linear interpolation (LM_{g+m}). For this last LM, weights were computed from the perplexity optimization on the EMEA development corpus, and vocabulary was fixed to 1 million words taking all the words of the in-domain corpus and the most frequent words from the out-of-domain corpora. Let us note that a high weight of 0.9 is associated with the medical LM despite its small size, which is explained by the great specificity of the medical domain.

Three Translation Models (TMs) incorporating a phrase table and a lexicalized reordering model are also built using MOSES: one (TM_g) from the out-of-domain data, one (TM_m) from the medical set and a last one (TM_{g+m}) from all the parallel corpora. For that purpose, bilingual data are aligned at the word level using the IBM 4 model (Och and Ney, 2003) with MGIZA++ (Gao and Vogel, 2008) . The score weights of a given TM and a selected LM are finally computed in each tested configuration using the Minimum Error Rate Training (MERT) method (Och, 2003) to optimize BLEU on the EMEA development corpus. To mix the information from the out and in-domain in TM_{g+m}, we resorted to the multiple translation tables option implemented into MOSES. With this feature, we can provide two translation tables to the decoder; the decoder first retrieves translation pairs from the in-domain phrase table, and resorts to the out-of-domain phrase-table as a fall-back.

3.3 SPE Systems

In order to build the SPE system for domain adaptation, we decide to translate the EMEA training corpus with each tested SMT system. Then, with the output of each system aligned with its translation reference, we build an SPE model using MOSES with default parameters. For the tuning process, we used the same in-domain development data as the SMT systems, this time with the SMT output aligned with its translation reference. Let us note that the weight optimization was repeated for each tested PBMT configuration.

4 Translating In-Domain Data

The first set of experiments deals with the translation of the domain specific, or in-domain, test corpus. The results are given in terms of BLEU scores in Table 2 with several uses of the previously described TMs and LMs. Pairwise comparisons between systems is made using approximate randomization as implemented in the evaluation tool FASTMTEVAL (Stroppa et al., 2007). These results indicate that the best configuration is $TM_{g+m}LM_{g+m}$, with a BLEU score of 47.3%. This score is not significantly higher (p-value=0.75) than the one obtained by $TM_{g+m}LM_m$ with an in-domain language model. These observations show that the specificity of the medical domain, including terminology and syntactic structures, cannot be improved by the introduction of out-of-domain data into the LM. For the translation model, however, the combination of the two phrase tables is the best configuration in the presented system comparison.

SMT system		% BLEU	p-value
TM_g	LM_g	29.9	0.002
TM_g	LM_{g+m}	38.2	0.002
TM_g	LM_m	39.2	0.002
com		44.9	0.007
TM_m	LM_m	46.4	0.001
TM_{g+m}	LM_m	47.2	0.75
TM_{g+m}	LM_{g+m}	47.3	

Table 2: BLEU scores of the different initial SMT systems when translating the test corpus from the medical domain.

The same conclusion about the importance of in-domain data can be derived from the results obtained with TM_g built on the sole out-of-domain data. A 10 points BLEU improvement is indeed obtained using LM_m instead of LM_g. Interpolating the two LMs introduces noise and decreases by 1 BLEU point the result obtained with LM_m only. Finally, let us note that the online system GOOGLE TRANSLATE has a BLEU score only 1.5 points lower than a PBMT system built using small-sized but highly relevant data.

5 Post-Editing Translations

After the translation step, SMT outputs are post-edited. Several SPE models are built from the translations of the EMEA training corpus generated by each SMT system. We decide to compute two scores: a first one for which all the sentences from the test corpus are post-edited, and a second one for which only sentences are post-edited if their sentence-level BLEU is improved (*oracle*). The computation of this *oracle* score relies on the reference translation and is done to estimate the potential of SPE.

5.1 Online System

The online translation tool already leads to good results in terms of BLEU score. The in-domain test corpus translated by the online system is post-edited by its SPE system. The results are shown in Table 3. Computating p-values to compare results before and after SPE exhibits a significant difference ($p = 0.001$ for BLEU and $p = 0.05$ for the *oracle* score).

System	% BLEU ($oracle$)
com	44.9
+ SPE_mLM_m	46.8 (53.3)
+ SPE_mLM_{g+m}	47.9 (53.5)

Table 3: BLEU scores of SPE on the online system output.

Two SPE systems are built with a different LM. With the medical LM (SPE_mLM_m), the BLEU score of the post-edited translation reaches 46.8%, around 2 points above the SMT output BLEU score. The *oracle* score indicates that more than 6 BLEU points can still be gained if the post-editing is only applied to the improvable subset of sentences from the test corpus. Introducing the out-of-domain LM with SPE_mLM_{g+m} leads to a BLEU score of 47.9%. The highest BLEU score obtained by an initial SMT (47.2% with the system $TM_{g+m}LM_m$) is already overtaken by this last SPE system jointly used with the com SMT system. Since the *oracle* scores indicate that the highest gain can be reached by the SPE system with the interpolated LM, we will focus on this configuration for our experiments on sentence selection described in Section 6.

5.2 Out-of-Domain PBMT System

This section describes the post-editing of out-of-domain PBMT system outputs, for which medical data are only employed to build LMs. For each LM used during the translation step, we evaluate the impact of the proposed SPE approach.

5.2.1 Out-of-Domain LM

The first evaluation of SPE on the out-of-domain PBMT system is done with TM_gLM_g relying only on out-of-domain data to build its statistical models. We introduce the in-domain data during the SPE step, in the SPE model, in the LM, or in both. The results are presented in Table 4. We can see

System	% BLEU ($oracle$)
TM_gLM_g	29.9
+ SPE_mLM_m	43.4 (44.2)
+ SPE_mLM_{g+m}	45.6 (47.0)

Table 4: BLEU scores of SPE on the out-of-domain PBMT system using an out-of-domain LM.

that introducing in-domain data during the post-editing step increases the BLEU score of the translated test corpus. From a baseline at 29.9% of BLEU, the SPE systems lead to an absolute improvement of 13.5 and 15.7 points depending on the SPE data configuration. Using the interpolated LMs for the SPE system shows the highest BLEU score, both with a *naive* application of SPE or for the *oracle* score. Let use note that the difference between SPE_mLM_m and SPE_mLM_{g+m} is statistically significant since it is associated with a p-value of 0.001. However, these results are lower than the BLEU score obtained by the specialized translation system (TM_mLM_m) presented in Table 2.

5.2.2 In-Domain LM

The second evaluation of SPE on the out-of-domain PBMT system concerns TM_gLM_m,

where in-domain data are introduced during the SMT process through the LM. The baseline is 39.2% of BLEU and the results presented in Table 5 show that 3.5 BLEU points are gained by the SPE step with a system built on medical data only. We performed the pairwise comparisons with BLEU and the *oracle* score and observed that SPE_mLM_m is statistically equivalent to SPE_mLM_{g+m} with $p > 0.1$ for both metrics. Again, these results are lower than the BLEU score obtained by the specialized translation system (TM_mLM_m) presented in Table 2.

System	% BLEU (*oracle*)
TM_gLM_m	39.2
+ SPE_mLM_m	42.7 (44.2)
+ SPE_mLM_{g+m}	42.5 (44.4)

Table 5: BLEU scores of SPE on the out-of-domain PBMT system using a medical LM.

5.3 In-Domain and Mixed PBMT Systems

After our experiments on the out-of-domain PBMT system using different LMs, we focus on the post-editing of in-domain PBMT system output. Two systems are studied here, one using only in-domain data (TM_mLM_m) and the other using both out and in-domain data ($TM_{g+m}LM_m$). For TM_mLM_m, the baseline BLEU score is 46.4% and none of the tested SPE configuration was able to increase this score. However, the *oracle* scores measured resp. at 47.4 % and 47.5 % with SPE_mLM_{g+m} and SPE_mLM_m show the potential improvement using SPE. This aspect motivates our sentence selection approach presented in Section 6.

As far as $TM_{g+m}LM_m$ is concerned, the use of the interpolated LM in the post-editing step (SPE_mLM_{g+m}) degrades the BLEU score by 0.8 point, while the use of the medical LM (SPE_mLM_m) does not statistically improve the baseline BLEU measured before SPE. For both configuration, the *oracle* score shows that a significant gain is still possible.

6 Selecting Sentences for Post-Editing

Post-editing selected sentences is motivated by the *oracle* scores measured in Section 5. We propose to build a classifier in order to partition sentences according to the possible BLEU gain with SPE. To train such a classifier, we use the medical development corpus and compute for each sentence its associated $\Delta BLEU$ score comparing BLEU before and after SPE. It is a binary classification task: if the $\Delta BLEU$ score is positive, i.e. SPE improves the sentence, the sentence is labelled Class 1; otherwise, the sentence is tagged with Class 2. Figure 1 illustrates the general architecture of our system.

The classifier used in our experiments is a Support Vector Machine (SVM) (Boser et al., 1992) based on a linear kernel. We use the implementation of *libSVM* (Chang and Lin, 2011) in the WEKA (Hall et al., 2009) environment (EL-Manzalawy and Honavar, 2005). The translated (by the MT system com) in-domain development set is used to build a sentence-level post-edition model. Each sentence of the training corpus is considered as a vector composed of n-grams ($n \in [1;3]$).

We decided to apply the classification method to the highest *oracle* score observed in Section 5, i.e. the com translation system jointly used with a SPE_mLM_{g+m} post-editing step. The *oracle* score for this configuration reaches 53.5%, while the *naive* application of SPE leads to a BLEU score of 47.9%. The test set translated by com is classified using SVM, where each sentence is associated with a normalized score for each of the two classes. Using the translation reference, we evaluate the classifier in terms of recall and precision. The recall reaches 79.5% and the precision 40.1%. In order to evaluate the gain in terms of BLEU on the whole test set, we decide to post-edit sentences according to their Class 1 scores given by the SVM. This score is the probability to improve BLEU at the sentence level. The evaluation can be repeated individually for each 0.1 score span (*is*, only the sentences in this exact range are post-edited) and then cumulated over consecutive spans (*cs*, all sentences above the threshold are post-edited). The results are displayed in Figure 2.

The cumulated span evaluation shows that post-editing the sentences above a prediction score of 0.8 reaches the highest BLEU score. With this configuration, 1 BLEU point is gained compared to the *naive* application of SPE (from 47.9% to 48.9% of BLEU). The amount of sentences in each class is increasing between 0.5 and 0.8. Only 60 sentences remain in Class 1 with a prediction score above 0.9. The amount of training sentences

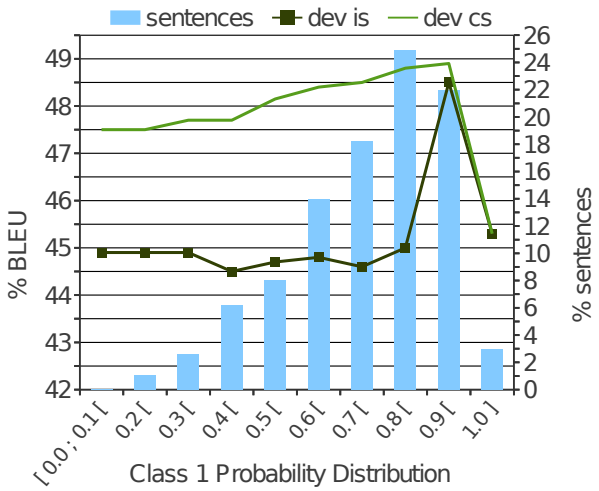

Figure 2: BLEU scores and amount of sentences classified in Class 1 for individual (*is*) and cumulated (*cs*) spans obtained on the test corpus.

in each class is an important aspect of the classifier accuracy. Figure 3 shows TER (Snover et al., 2006) and inverted BLEU scores of Class 1 sentences with a classification score over 0.8, before and after post-editing.

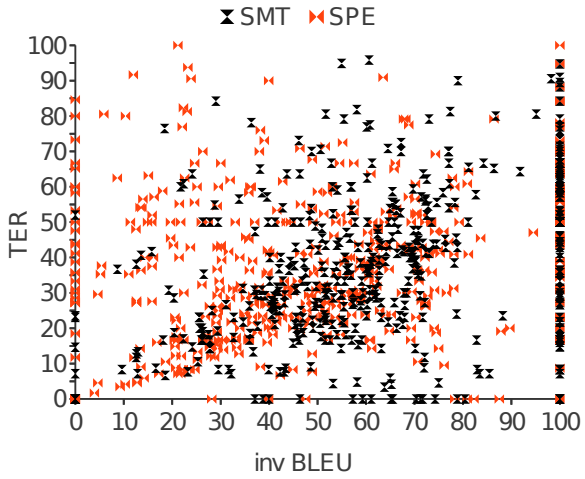

Figure 3: TER and *inverted* BLEU sentences distribution measured on the test corpus when Class 1 probability is over 0.8.

It clearly appears that there are more post-edited than translated sentences with a 100% BLEU score (0% inverted BLEU): resp. 47 and 11 sentences. Also among the 109 translated sentences with a 0% BLEU score, only half remains at this level after post-editing. The evaluation on the test set shows a general improvement using both metrics, as detailed in Table 6. These final results present the possible gain in terms of translation quality with SPE and a classification approach. The comparison between the SPE systems with and without classification shows that the combination of SPE and SVM is better than the *naive* application of SPE with $p = 0.004$.

	SMT	+ SPE	+ SVM
TER	42.3	40.4	39.7
BLEU	44.9	47.9	48.9

Table 6: TER and BLEU scores on the test set after translation, post-editing and classification (with $p(Class1) >= 0.8$).

7 Conclusion and Future Work

In this paper, we have presented a phrase-based post-editing approach for specific domain adaptation. Our experiments show that an out-of-domain translation system can be adapted *a posteriori* through a *naive* application of the proposed SPE approach. *Oracle* scores indicate that gains in terms of BLEU score are still possible, even with a PBMT system built on in-domain data and without introducing new data during the post-editing step. The highest BLEU score is obtained using GOOGLE TRANSLATE combined with an SPE system ($SPE_m LM_{g+m}$) and a classification step. Compared to the baseline, the BLEU score is increased by 4 BLEU points. Compared to the best PBMT system ($TM_{g+m} LM_{g+m}$) with 47.3% of BLEU, the score is increased by 1.6 BLEU points (with $p = 0.001$). In a future work, other metrics will be used to measure the translation quality at the sentence level. We also want to introduce more features into the classifier training set based on quality estimation techniques for our sentence selection approach, in order to better fill the gap between the current BLEU and the *oracle* score.

References

Allen, J. and C. Hogan. 2000. Toward the development of a post editing module for raw machine translation output: A controlled language perspective. In *CLAW*, pages 62–71.

Béchara, H., Y. Ma, and J. van Genabith. 2011. Statistical post-editing for a statistical MT system. In *MT Summit XIII*, pages 308–315.

Boser, B.E., I.M. Guyon, and V.N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *5th annual workshop on Computational learning theory*, pages 144–152.

Chang, Chih-Chung and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

de Ilarraza, A.D., G. Labaka, and K. Sarasola. 2008. Statistical postediting: A valuable method in domain adaptation of RBMT systems for less-resourced languages. In *MATMT*, pages 35–40.

Dugast, L., J. Senellart, and P. Koehn. 2007. Statistical post-editing on Systran's rule-based translation system. In *WMT*, pages 220–223.

Dugast, L., J. Senellart, and P. Koehn. 2009. Statistical post editing and dictionary extraction: Systran/Edinburgh submissions for ACL-WMT2009. In *WMT*, pages 110–114.

EL-Manzalawy, Y. and V. Honavar, 2005. *WLSVM: Integrating LibSVM into Weka Environment*. Software available at `http://www.cs.iastate.edu/~yasser/wlsvm`.

Gao, Q. and S. Vogel. 2008. Parallel implementations of word alignment tool. In *ACL Workshop: Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Isabelle, P., C. Goutte, and M. Simard. 2007. Domain adaptation of MT systems through automatic post-editing. In *MT Summit XI*, pages 255–261.

Knight, K. and I. Chander. 1994. Automated postediting of documents. In *NCAI*, pages 779–779.

Koehn, P., F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *NAACL-HLT*, volume 1, pages 48–54.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*, pages 177–180.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*, volume 5, pages 79–86.

Marcu, D. and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *EMNLP*, volume 10, pages 133–139.

Och, F.J. and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Och, F.J. 2003. Minimum error rate training in statistical machine translation. In *ACL*, volume 1, pages 160–167.

Oflazer, K. and I.D. El-Kahlout. 2007. Exploring different representational units in English-to-Turkish statistical machine translation. In *WMT*, pages 25–32.

Papineni, K., S. Roukos, T. Ward, and W.J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Rafalovitch, A. and R. Dale. 2009. United Nations general assembly resolutions: A six-language parallel corpus. *MT Summit XII*, pages 292–299.

Sadat, F., J.H. Johnson, A. Agbago, G. Foster, R. Kuhn, J. Martin, and A. Tikuisis. 2005. PORTAGE: A phrase-based machine translation system. In *The ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*.

Sager, J.C., D. Dungworth, and P.F. McDonald. 1980. English special languages: Principles and practice in science and technology.

Simard, M., C. Goutte, and P. Isabelle. 2007a. Statistical phrase-based post-editing. In *NAACL-HLT*, pages 508,515.

Simard, M., N. Ueffing, P. Isabelle, and R. Kuhn. 2007b. Rule-based translation with statistical phrase-based post-editing. In *WMT*, pages 203–206.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *AMTA*, pages 223–231.

Stolcke, A. 2002. SRILM—an extensible language modeling toolkit. In *InterSpeech*, volume 2, pages 901–904.

Stroppa, N., K. OwczarBézak, and A. Way. 2007. A cluster-based representation for multi-system MT evaluation. In *TMI*, pages 221–230.

Suzuki, H. 2011. Automatic post-editing based on SMT and its selective application by sentence-level automatic quality evaluation. In *MT Summit XIII*, pages 156–163.

Tiedemann, J. 2009. News from OPUS—a collection of multilingual parallel corpora with tools and interfaces. In *RANLP*, volume V, pages 237–248.

Zens, R., F. Och, and H. Ney. 2002. Phrase-based statistical machine translation. *KI 2002: Advances in Artificial Intelligence*, pages 35–56.