Domain Adaptation Techniques for Machine Translation and Their Evaluation in a Real-World Setting

Baskaran Sankaran[1,⋆], Majid Razmara[1,⋆], Atefeh Farzindar[2], Wael Khreich[2],
Fred Popowich[1], and Anoop Sarkar[1]

[1] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
{bsa33,mra44,popowich,anoop}@sfu.ca
[2] NLP Technologies Inc., 52 Le Royer, Montreal, QC, Canada
{farzindar,wael}@nlptechnologies.ca

Abstract. Statistical Machine Translation (SMT) is currently used in real-time and commercial settings to quickly produce initial translations for a document which can later be edited by a human. The SMT models specialized for one domain often perform poorly when applied to other domains. The typical assumption that both training and testing data are drawn from the same distribution no longer applies. This paper evaluates domain adaptation techniques for SMT systems in the context of end-user feedback in a real world application. We present our experiments using two adaptive techniques, one relying on *log-linear models* and the other using *mixture models*. We describe our experimental results on legal and government data, and present the human evaluation effort for post-editing in addition to traditional automated scoring techniques (BLEU scores). The human effort is based primarily on the amount of time and number of edits required by a professional post-editor to improve the quality of machine-generated translations to meet industry standards. The experimental results in this paper show that the domain adaptation techniques can yield a significant increase in BLEU score (up to four points) and a significant reduction in post-editing time of about one second per word.

Keywords: Statistical Machine Translation, Machine Learning, Domain Adaptation, Model Adaptation, Human Evaluation.

1 Introduction

Statistical Machine Translation (SMT) uses machine learning methods to train high quality translation systems from large corpora, consisting of parallel aligned sentences in source and target languages. Good performance of SMT systems depends on the availability of a sufficient amount of high-quality parallel corpora. When provided with a large amount of parallel training sentences from a specific domain, SMT models can produce accurate translations for previously unseen sentences that belong to the same domain. In fact, SMT models can approximate the underlying data distribution from training corpora and generalize to novel test data drawn from the *same* distribution.

[⋆] B. Sankaran and M. Razmara contributed equally to this work.

L. Kosseim and D. Inkpen (Eds.): Canadian AI 2012, LNAI 7310, pp. 158–169, 2012.

In practice, large corpora are typically available for general domains. For instance, the documents translated by international organizations such as the European Parliament, United Nations and Canadian Hansard are among the largest parallel corpora currently available. These documents are usually collected from diverse areas, and hence belong to various domains. SMT models specialized for one domain would perform poorly when applied to other domains; the typical assumption that both training and testing data are drawn from the same distribution is no longer valid. For instance, variations in language vocabulary, writing style or grammar yield different distributions across domains. In practice, collecting and manually labelling representative training corpora for specific domains could be prohibitively expensive. Therefore, it would be more efficient to adapt SMT models trained on a general domain to specific domains, than to train and maintain specific models for each domain.

Domain adaptation techniques allow SMT models to generalize from a source domain with abundant data to a different target domain with limited data. These techniques vary from supervised [5,8,12,4] to semi-supervised [20] and unsupervised [21,2] domain adaptation. Supervised domain adaptation techniques assume that limited parallel data is available from the target domain, while unsupervised domain adaptation techniques rely solely on target monolingual data. Supervised domain adaptation techniques for phrase-based SMT systems includes manipulation of source and target corpora [5], and adaptation of language and translation models using log-linear and linear mixture models [8,12,4]. Other techniques including system combination approaches have also been used in domain adaptation [10,6,16].

Domain adaptation is of interest for NLP Technologies and other companies providing translation services. While NLP Technologies has assembled large bilingual corpora (over 1.6 million sentence pairs) for the legal domain, there are continuous requests for new specific domains for which there are limited amounts of parallel sentences. Adaption of current SMT systems to these domains would decrease the amount of time and costs required for translation. Furthermore, translation quality of a previously unseen test set would improve because human translators will have more time to focus on post-editing tasks (e.g., contextual accuracy). This high quality post-edited text could also be used to revise the SMT models and also increase the accuracy of the SMT systems over time.

In this paper, we examine the adaptation of a general SMT system trained on NLP Technologies' Legal corpora (NL) to two specific legal domains, each with a limited amount of parallel sentences. One of the target domains focuses on English-French translation of legislative documents for the Indian and Northern Affairs (IA), while the other focuses on French-English translation of the judgments of the Human Rights (HR) commission in Quebec. All experiments are conducted using the phrase-based SMT system PORTAGE developed at the National Research Council of Canada (NRC) [17].

The performance of the domain adaptation techniques is evaluated using the BLEU evaluation metric [15], which measures the translation quality by computing and combining the precision of different n-grams in the system output to that of (possibly multiple) reference translations. Since these solutions will be deployed in an operational environment, the outputs from each adapted system are also evaluated by professional human translators. The human evaluation is based primarily on the amount of time required by a professional post-editor to improve the quality of machine-generated translations to industry standards. In addition, the post-editing effort is also measured by the Human-targeted Translation Edit Rate (HTER) [18].

Section 2 of this paper describes the approach taken by a traditional SMT system, which we refer to as the baseline system because it does *not* incorporate any domain adaptation method. Section 3 presents the domain adaptation techniques for SMT employed in this work. In Section 4, the experimental results in terms of BLEU scores and human evaluation metrics (post-editing time and HTER) are presented and discussed. Finally, the conclusions and future work are presented in section 5.

2 Baseline Translation System

The goal of a machine translation system is to translate a source language sentence $s = s_1 s_2 \cdots s_J$ into a target language sentence $t = t_1 t_2 \cdots t_I$. We use phrase-based statistical machine translation [11] in this work, which can be represented in a log-linear framework [14] consisting of a set of feature functions h_m, as below:

$$P(t|s) = \frac{1}{Z} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(t,s)\right)$$

where λ_m are the set of weights corresponding to M feature functions and Z a normalization factor. This log-linear formulation leads to the following approximation:

$$\hat{t} = \operatorname*{argmax}_{t \in T} \left[\sum_{m=1}^{M} \lambda_m h_m(t,s)\right] \tag{1}$$

The weights (λ_m) are estimated iteratively to maximize the likelihood of the training data or trained on a development set to directly minimize a translation error criterion [13]. The number of function features (M) is only limited by the computational power available and the time allocated for optimization. A typical set of feature functions include:

- Phrase probabilities in both translation directions $P(t|s)$ and $P(s|t)$, which specify alternative translation candidates and their probabilities for each source and target phrase.
- Lexical probabilities in both translation directions $P_l(t|s)$ and $P_l(s|t)$, which indicate how well individual words translate to each other.
- Language model $P(t)$, which provides the likelihood of a candidate translation being a proper sentence of the target language.
- Word penalty $W(t)$, which calibrates the output length by penalizing very long or very short target sentences.
- Distortion model $d(s,t)$, which reorders phrases and favors translation candidates with proper phrase order for the target language.

In our experiments, each of the phrase and lexical probabilities are computed using IBM model 2 and HMM alignments, which provides 11 features for the training and optimization of the baseline SMT systems. This is achieved by using the phrase-based SMT system PORTAGE [17], which provides a platform that computes and combines these features using the log-linear framework that was shown in (1). The weights are determined according to the minimum error rate training (MERT) algorithm [13] using the BLEU metric [15] as an optimization criterion. The MERT algorithm finds optimal feature weights by performing a line search for each parameter (independently) to maximize the BLEU score.

3 Domain Adaptation

When designing SMT systems for a particular domain, it is usually assumed that both training and testing data are drawn from the same distribution D. In domain adaptation however, the objective is to adapt a SMT system trained on source domain (i.e. out-of-domain) with a distribution D_{out}, to a target domain (i.e. in-domain) with a different distribution D_{in}. The supervised domain adaptation techniques we considered involved log-linear and mixture models [8].

With the log-linear approach, data from the different domains D_{out} and D_{in} are treated as separate features which are then combined in a log-linear framework by using distinct feature weights each of which is tuned by the MERT. While simpler, this has the disadvantage of *increasing* the number of features in the framework.

In this approach, the models of the out-of-domain data D_{out} and the in-domain data D_{in} are treated as distinct features in the global log-linear model apart from the regular features of the baseline system listed in Section 2. The result is a total of 20 features, including 9 (8 for the translation model and 1 for the language model) additional features that need to be tuned by MERT.

As observed by Chiang et al. [3], the number of features that can be reliably optimized by MERT can be as low as 15. Earlier experiments [3] have shown MERT to be inefficient in handling a large number of features, while optimizing the feature weights. However on the positive side, treating them as distinct features allows their weights to be directly optimized by the MERT for the BLEU evaluation metric. Furthermore, this system can be retrained more quickly if additional in-domain data is available.

Using the *mixture model* approach, also referred to as *linear mixtures* [17], we can *mix* the models pertaining to different domains into a single model by a weighted combination of the components. In a mixture model, the conditional probability of each phrase-pair would be the weighted sum of the conditional probabilities of the phrase-pair in two phrase tables.

$$P(t|s) = wP_{in}(t|s) + (1 - w)P_{out}(t|s)$$

where w is the interpolation weight and $0 \leq w \leq 1$. So, a phrase observed in both D_{out} and D_{in} will get a higher feature value than a phrase seen in either domain.

The mixture model approach has been shown to yield better results in a particular setting of combining the mixture components in a linear way [8] and has the advantage of not increasing the total number of features. However, the downside of this approach is that the mixture components do not participate in the global log-linear model directly and thus it is difficult to set the weights for mixture components [8].

In the experiments reported in the next section, the NL corpus is used as D_{out}. Language models pertaining to different domains are combined using the log-linear framework, while we experiment with the log-linear and mixture approaches for combining the translation models. First two phrase tables are learned, one based on D_{out} and the other for D_{in}. Second, these two phrase tables are merged using the mixture models approach with different interpolation weights. The resulting phrase table then is used in a log-linear framework to assign an optimal weight as in PORTAGE original work flow. Section 4.1 summarizes the result of our system using this approach when applying different weights for interpolation.

4 Experimental Results

Table 1 provides a summary of the data sets used for the training, optimization and testing of each SMT system for two target domains, one related to Human Rights (HR) data, and the other related to Indian Affairs (IA) data. As shown in Table 1, the number of parallel sentences from the source domain (D_{out}), provided by NLP Technologies' Legal corpora (NL), is larger by about two orders of magnitude than that of the HR and the IA domains.

Table 1. Corpus statistics for source and target domains

	Number of sentence pairs			Average sentence length					
Domain				English			French		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
NL	1631153	-	-	19.4	-	-	23.8	-	-
HR	16444	1000	1000	20.2	16.6	19.7	20.9	18.0	22.1
IA	21037	1000	1000	6.7	6.5	6.3	8.5	8.2	7.8

Three different baseline systems are built for performance comparison with the domain adaptation techniques. The first system (Baseline 1) is trained and optimized using data from NL domain only. Parameter optimization is performed on 1000 sentence pairs from NL data set that were not included in training (hold-out validation). Baseline 2 is trained on NL training set and optimized on the development set from the target domains (HR or IA). The third baseline is trained on the concatenation of the source and target training sets and optimized using the development set of the target domains. For the Human Rights domain, for instance, Baseline 3 is trained on (NL ∪ HR) training set and optimized on HR development set. All systems are evaluated on the testing sets from the target domains. Note that the results are based on a random split in training, development and test sets.

To evaluate the domain adaptation techniques, we start by providing BLEU score evaluations, comparing the results on the two different data sets to those obtained with the baseline approaches. The best of the domain adaptation techniques are then incorporated into the *Adaptive TRANSLI* system (Adaptive Translation of Legal Information), which is then compared to the current machine translation system employed at NLP

Technologies. This comparison involves undertaking a human evaluation of the translation quality of the two systems.

4.1 BLEU Score Evaluation of Domain Adaptation Techniques

Table 2 presents the performance evaluation of the domain adaptation techniques as measured by BLEU for both domains – HR and IA – compared to that of the baselines. As shown in Table 2, the log-linear model has achieved a slight improvement over Baseline 3 and linear mixture model for the HR domain adaptation. By contrast, the level of performance of the mixture model is substantially higher than that of the log-linear model for the IA domain adaptation, with a BLEU score improvement of 2.75 over Baseline 3. In practice, one adaptation technique can be prefered to the other based on their performance on a set.

Table 2. BLEU score results for adaptation to Human Rights and Indian Affairs domains

System	HR	IA
	Fr → En	En → Fr
Baseline 1	40.54	23.60
Baseline 2	38.35	25.20
Baseline 3	41.91	26.34
Log-linear	**41.97**	25.22
Mixture	41.33	**29.09**

Table 3 captures the BLEU score variations for the mixture model, under different combinations of mixture weights for in-domain and out-of-domain models. The weights of different feature functions (e.g. phrase table, language models, etc.) are tuned using *minimum error rate training (MERT)* based on a held-out set of *HR* data.

Table 3. Mixture-Model Results for *NL* and *HR* data

Weights		Fr → En	En → Fr
w_{NL}	w_{HR}		
0.5	0.5	41.21	36.50
0.4	0.6	41.09	36.49
0.3	0.7	41.33	**37.11**
0.2	0.8	41.45	36.88
0.1	0.9	**41.82**	36.84

Table 4 similarly shows the results when using the *NL* corpus for out-of-domain and *IA* as in-domain data. Here, the variance in the BLEU scores is higher than that found in the HR domain. We attribute this to a greater degree of noise in the data, particularly in the segmentation and alignment of sentences. Less noisy data training data would improve the performance.

The results show that the $(0.3, 0.7)$ weighting yields the highest BLEU scores for English to French translation of HR sentences, and for French to English translation

Table 4. Mixture-Model Results for *NL* and *IA* Data

Weights		Fr → En	En → Fr
w_{NL}	w_{IA}		
0.5	0.5	28.02	**29.09**
0.4	0.6	27.91	28.83
0.3	0.7	**28.69**	28.60
0.2	0.8	27.21	23.82

in the IA domain. For English to French translation in the IA domain, the $(0.5, 0.5)$ weighting provided the best BLEU score, but only approximately 0.5 higher than the $(0.3, 0.7)$ weighting. However, the highest BLEU score in the French-to-English direction for the HR domain is achieved when a more skewed weighting $(0.1, 0.9)$ is used, but again only a 0.5 improvement over the $(0.3, 0.7)$ weighting.

In this work we only use data from the legal domain for training while adapting to specific sub-domains. However, the domain adaptation MT system could also take into consideration the large parallel data that is not in the legal domain and this could possibly result in further improvements. This is trivially possible in the mixture model since we would simply add in an extra component in the mixture without increasing decoding complexity. However, more sophisticated means are required to adapt the log-linear model with multiple domains in the training data. Some of the authors are currently focussing on this problem.

4.2 Human Evaluation: Methods and Results

In previous work, Farzindar et al. [7,9,19] have experimented with various human evaluation techniques to assess the quality and fidelity[1] of SMT output. These techniques include the Levenshtein edit distance applied to space-separated tokens (any sequence of contiguous non-space characters) that differ between the SMT output and the output revised by a human post-editor. The number of operations, which measures the number of *consecutive* insertion, deletion, and substitution operations, required by a post-editor to revise the SMT translation in the context of federal court judgments. This measure is different from the edit distance, since it approximates the number of cut and paste operations needed to revise an SMT translation. For example, by substituting five consecutive words the edit distance would be five, whereas the number of operations is equal to one.

The authors have also proposed an approach to estimate the post-editing effort of translations produced by SMT systems at the sentence level, to prevent post-editors from revising low quality translations, which could be more time consuming than a direct translation [19]. Computing the post-editing effort is based on various features reflecting the difficulty of translating the source sentence and the discrepancies between the source and translation sentences [19], and measured using the Human-targeted Translation Edit Rate (HTER) [18]. The HTER is defined as the minimum number of

[1] The fidelity measures the amount of information (semantic content) properly transferred from the source sentence to the target sentence.

edits required to change an SMT output to match the reference, normalized by the length of the reference. Edits include insertion, deletion and substitution of single words, as any standard edit distance metric, as well as shifts of word sequences.

$$HTER = \frac{\#edits}{\#reference\text{-}words}$$

In this paper, the evaluation is based primarily on the amount of time required by a professional post-editor to improve the quality of machine-generated translations to industry standards. Since translation quality will always be ensured by post-editors, in professional translation companies, the objective is to evaluate the reduction in post-editing time achieved by the adapted SMT systems. Less post-editing time implies superior machine translation quality. In fact, time is a critical factor for translation companies because translators are typically paid per word count or per hour. In addition, the post-editing effort measured by the HTER is also provided for reference.

The current human evaluation experiments involve the translation of two documents from each domain (HR1, HR2, IA1 and IA2), where each document contains about 400 words (see Table 5 for details). None of the documents were included in any corpus previously-used for training, tuning or testing. These documents are translated by two different systems. The Adaptive TRANSLI system incorporates the best adaptive models described in the previous section, either the log-linear or mixture models depending on the domain and with best interpolation weights in case of mixture models. The current operational SMT system at NLP Technologies is the second system, which is used for comparison purposes. The source and machine translated outputs are then integrated into the post-editing tool (PET) developed by Aziz et al. for assessing machine translation [1]. PET allows to measure the time and HTER values required to post-edit each sentence, and offers other interesting features such as recording all changes made during the post-editing process.

Four professional post-editors from NLP Technologies were asked to post-edit the machine translated outputs to meet industry standards. The post-editors were selected to revise the translations in their native languages. They had no knowledge whether the documents were translated by the Adaptive TRANSLI or NLP current system. For unbiased evaluation, the order in which the machine-translated documents were presented to post-editors was randomized, and an interval of one week was left between each revision of the different translations of the same document.

Table 6 presents the average post-editing time (in seconds) per word as well as per sentence and the average HTER values per sentence for each post-edited document and system. As shown in Table 6, the average time per word required by all post-editors to

Table 5. Statistics about documents selected for human evaluation

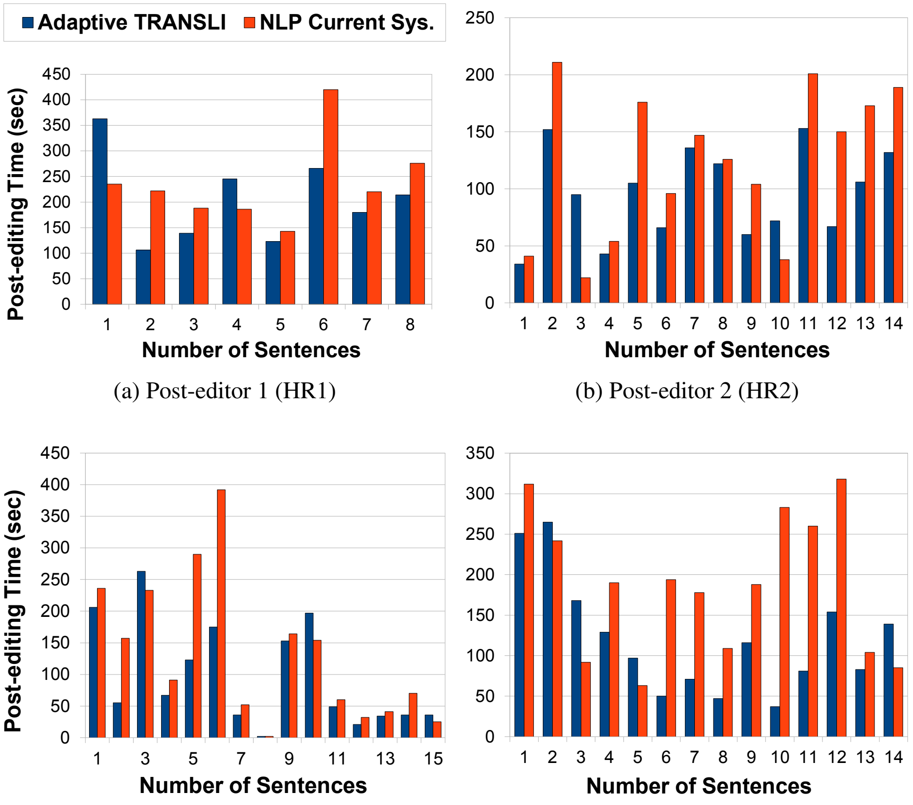
Document	#Words	#Sentences	Avg. sentence length	Translation
HR1	449	8	56.1	Fr→En
HR2	443	14	31.6	Fr→ En
IA1	407	15	27.1	En→Fr
IA2	392	14	28	En→Fr

Table 6. Average post-editing time and HTER values during human evaluation

System	Doc	Avg time per word (sec)	Avg time per sentence	Avg HTER per sentence
Adaptive TRANSLI	HR1	**3.6**	**204** ± 86	0.61
	HR2	**3.0**	**95** ± 39	0.49
	IA1	**3.6**	**97** ± 80	**0.28**
	IA2	**4.3**	**120** ± 70	**0.32**
NLP Current Sys.	HR1	4.2	238 ± 81	**0.54**
	HR2	3.9	122 ± 65	**0.38**
	IA1	4.9	132 ± 112	0.29
	IA2	6.7	187 ± 86	0.49

revise the output of Adaptive TRANSLI is on average overall lower than that of NLP current system by about one second. The table also shows that the average time spent to post-edit a sentence translated by Adaptive TRANSLI is significantly less than that of a sentence translated by the current system. The adaptive system could reduce the post-editing time by (on average) 66 seconds in 78% of the sentences. In the rest of the sentences (22%), the average post-editing time was increased by 56 seconds per sentence. This can be also seen in more detail in Figure 1, where the time required to post-edit the output of each system is illustrated for each sentence. With a daily translation capacity of $10,000$ words for instance, an average of one second reduction in post-editing time per word according to Adaptive TRANSLI would save about 2.8 hours. Therefore, the company would be able to process larger number of requests on daily basis. In addition, the human translators would have more time to focus on contextual accuracy and overall quality of translations.

The numbers of edits per sentence captured in the HTER values are shown to decrease with the Adaptive TRANSLI models for the IA domain. However, for the HR domain the HTER values produced by the Adaptive TRANSLI models are higher than that of NLP current system, as shown in Table 6. This is mainly caused by some general French phrases that occurred in the Human Rights documents, such as "mise en cause" and "éléments suivants" and remained untranslated with the adaptive TRANSLI, while they have been translated by NLP current system. The Post-editors needed to perform additional number of edits to translate the phrases that remained untranslated with the adaptive TRANSLI, which explains the high values of HTER for the French-English translation of the Human Rights documents. In fact, the current system employed at NLP is trained on large and diverse corpora in addition to the corpora from the legal domain, and hence it is able to translate some general expressions better than the Adaptive TRANSLI. However, the overall time required to post-edit a document from the Human Rights domain translated by the Adaptive TRANSLI remained lower than that of NLP current system, which demonstrates the high level of precision provided by the adaptive system.



(a) Post-editor 1 (HR1)

(b) Post-editor 2 (HR2)

(c) Post-editor 3 (IA1)

(d) Post-editor 4 (IA2)

Fig. 1. A comparison of post-editing time (in seconds) for each sentence. Lower post-editing time implies superior translation quality.

5 Conclusion and Future Work

This paper presents an approach for adapting machine translations systems to new domains along with an evaluation of the domain adaptation techniques both in terms of traditional automatic evaluation metrics and human evaluations in a real-world setting. Since there is a well established need for translations involving new specific domains with limited amounts of parallel sentences, the adaption of current SMT systems to new domains would reduce the translation time and efforts while maintaining translation quality. The Adaptive TRANSLI system, comprising two domain adaptation techniques based on log-linear models and mixture models, has been used in conjunction with the general PORTAGE SMT system on different legal domains – the Human Rights and the Indian and Northern Affairs.

Our experiments on log-linear and mixture models demonstrate the mixed strengths for both approaches and our results are similar to those of [8]. Additionally, the mixture model has the advantage that a large parallel corpus that is not in the legal domain could be used easily to further improve the translation quality. However, more sophisticated means are required to adapt the log-linear model with multiple domains in the training data and some of the authors in this work are currently focussing on this problem.

The results of automatic evaluation have shown that the adaptive techniques can yield a significant increase in the BLEU score (up to four points) over that of the baseline (with no adaptation). Human-evaluation results have shown a significant reduction in post-editing time of about one second per word, which would save about three hours daily in a production environment with a translation capacity of $10,000$ words per day.

The cleaning of the *new domain* training data by a combination of automatic methods and manual verification could further improve translation quality of a previously unseen test set. This could be done in conjunction with the real-world translation process used by companies providing translation services to improve the overall translation process. NLP Technologies Inc. is investigating the integration of the adaptive translation methods into its translation environment tools to reduce the post-editing time and effort at the sentence level, and allow the post-editors to focus further on overall translation quality. An interesting future extension to this work would consist of developing and implementing incremental and active learning techniques to interactively integrate post-editors' feedback into the SMT system during operations.

Acknowledgements. This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC), Canada through an Engage grant. We gratefully acknowledge the National Research Council (NRC), Canada for providing the Portage (version 1.4.2) machine translation system through the PortageShared technology transfer to Canadian universities. Special thanks to Lucia Specia (University of Sheffield) and to Wilker F. Aziz (University of Wolverhampton) for providing the post-editing tool.

References

1. Aziz, W., de Sousa, S.C.M., Specia, L.: PET: a tool for post-editing and assessing machine translation. In: The 8th International Conference on Language Resources and Evaluation (LREC 2012), Instanbul, Turkey (2012)
2. Bertoldi, N., Federico, M.: Domain adaptation for statistical machine translation with monolingual resources. In: Proceedings of the 4th Workshop on Statistical Machine Translation, StatMT 2009, pp. 182–189. Association for Computational Linguistics, USA (2009)
3. Chiang, D., Marton, Y., Resnik, P.: Online large-margin training of syntactic and structural translation features. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. ACL (2008)
4. Civera, J., Juan, A.: Domain adaptation in statistical machine translation with mixture modelling. In: Proceedings of the Second Workshop on Statistical Machine Translation, StatMT 2007, pp. 177–180. Association for Computational Linguistics, Stroudsburg (2007)
5. Daumé III, H.: Frustratingly easy domain adaptation. In: Conference of the Association for Computational Linguistics. ACL, Prague (2007)
6. DeNero, J., Kumar, S., Chelba, C., Och, F.: Model combination for machine translation. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT 2010, pp. 975–983. Association for Computational Linguistics, Stroudsburg (2010)

7. Farzindar, A., Lapalme, G.: Machine Translation of Legal Information and Its Evaluation. In: Gao, Y., Japkowicz, N. (eds.) AI 2009. LNCS, vol. 5549, pp. 64–73. Springer, Heidelberg (2009)
8. Foster, G., Kuhn, R.: Mixture-model adaptation for SMT. In: Proceedings of the Second Workshop on Statistical Machine Translation. ACL (2007)
9. Gotti, F., Farzindar, A., Lapalme, G., Macklovitch, E.: Automatic translation of court judgments. In: AMTA 2008 The Eighth Conference of the Association for Machine Translation in the Americas, pp. 1–10. Waikiki, Hawai'i (2008)
10. Hildebrand, A.S., Vogel, S.: CMU system combination for WMT 2009. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, StatMT 2009, pp. 47–50. Association for Computational Linguistics, Stroudsburg (2009)
11. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 48–54. Association for Computational Linguistics (2003)
12. Koehn, P., Schroeder, J.: Experiments in domain adaptation for statistical machine translation. In: Proceedings of the Second Workshop on Statistical Machine Translation, StatMT 2007, pp. 224–227. Association for Computational Linguistics, Stroudsburg (2007)
13. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the Association of Computational Linguistics. ACL (2003)
14. Och, F.J., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL 2002, pp. 295–302. ACL, Stroudsburg (2002)
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the Association for Computational Linguistics, pp. 311–318. ACL (2002)
16. Razmara, M., Foster, G., Sankaran, B., Sarkar, A.: Mixing multiple translation models in statistical machine translation. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Jeju (2012)
17. Sadat, F., Johnson, H., Agbago, A., Foster, G., Martin, J., Tikuisis, A.: Portage: A phrase-based machine translation system. In: Proceedings of the ACL Worskhop on Building and Using Parallel Texts. ACL (2005)
18. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas, pp. 223–231 (2006)
19. Specia, L., Farzindar, A.: Estimating machine translation post-editing effort with HTER. In: AMTA 2010 Workshop, Bringing MT to the User: MT Research and the Translation Industry. The 9th Conference of the Association for Machine Translation in the Americas (2010)
20. Ueffing, N., Haffari, G., Sarkar, A.: Semi-supervised model adaptation for statistical machine translation. Machine Translation 21(2), 77–94 (2007)
21. Wu, H., Wang, H., Zong, C.: Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In: Proceedings of the 22nd International Conference on Computational Linguistics, COLING 2008, USA, pp. 993–1000 (2008)