Evaluation of Domain Adaptation Techniques for TRANSLI in a Real-World Environment

Atefeh Farzindar and Wael Khreich
NLP Technologies Inc., 52 Le Royer, Montreal, QC, Canada
{farzindar, Wael}@nlptechnologies.ca

Abstract

Statistical Machine Translation (SMT) systems specialized for one domain often perform poorly when applied to other domains. Domain adaptation techniques allow SMT models trained from a source domain with abundant data to accommodate different target domains with limited data. This paper evaluates the performance of two adaptive techniques based on log-linear and mixture models on data from the legal domain in real-world settings. Performance evaluation includes post-editing time and effort required by a professional post-editor to improve the quality of machine-generated translations to meet industry standards, as well as traditional automated scoring techniques (BLEU scores). Results indicates that the domain adaptation techniques can yield a significant increase in BLEU score (up to three points) and a significant reduction in post-editing time of about one second per word in an operational environment.

1 Introduction

Statistical Machine Translation (SMT) systems specialized for one domain would perform poorly when applied to other domains. In fact, the typical assumption that both training and testing data are drawn from the same distribution is no longer valid. Variations in language vocabulary, writing style or grammar yield different distributions across domains. In practice, the collection and alignment of representative training corpora for specific domains could be prohibitively expensive. Therefore, it is 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. Domain adaptation is of interest for NLP Technologies and other companies providing translation services, since there are continuous requests for translation of new specific domains with limited amounts of parallel sentences. Adaption of current SMT systems to these domains would decrease the amount of time and costs required for translation, and hence provide more time for human translators to focus on post-editing tasks (e.g., contextual accuracy).

In this paper, we examine the adaptation of a general SMT system (Farzindar, 2009) trained on NLP Technologies' Legal corpora (NL) to two specific legal domains with 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 (see Table 1). We focus on supervised domain adaptation techniques for phrase-based SMT systems, which include adaptation of language and translation models using log-linear and mixture models (Foster and Kuhn, 2007). All experiments are conducted using the PORTAGE system developed at the National Research Council of Canada (Sadat et al., 2005).

The performance of each domain adaptation technique is first evaluated on both HR and IA domains, using the BLEU score (Papineni et al., 2002). The outputs of the best performing technique on each domain are then evaluated by NLP Technologies' post-editors in terms of post-editing time and effort.

2 Domain Adaptation

Phrase-based SMTs typically follow the log-linear framework (Koehn et al., 2003) to translate a source sentence s into a target sentence t:

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

where λ_m are the set of weights corresponding to M feature functions h_m and Z is a normalization factor. A typical set of feature functions include phrase-pair probabilities and lexical weighting in both translation directions, language model, word penalty, and distortion model. In our experiments, each of the phrase and lexical probabilities are computed using IBM model 2 and HMM3 alignments, which provides 11 features for the training and optimization of the baseline SMT systems using PORTAGE system (Sadat et al., 2005). The weights λ_m are determined according to the minimum error rate training (MERT) algorithm (Och, 2003), which performs a line search for each parameter (independently) to maximize the BLEU score (Papineni et al., 2002).

Domain adaptation aims to adapt a SMT system trained on source domain with a distribution D_S, to a target domain with a different distribution D_T. Supervised domain adaptation techniques include log-linear and mixture models (Foster and Kuhn, 2007).

With the *log-linear approach*, distinct feature functions are computed separately for each domain D_S and D_T, and then combined in a log-linear framework (1). This provides $M = 20$ features (9 additional features over those of the baseline) for optimization with MERT algorithm. However, with increasing number of features MERT optimization may become inefficient (Chiang et al., 2008).

In the *mixture model approach* different SMT models are trained for each domain, and then merged by a weighted combination of the components. For instance, the phrase-pair probabilities form each domain (ϕ_s and ϕ_t) are combined by:

$$\phi(t|s) = \lambda \phi_s(t|s) + (1 - \lambda) \phi_t(t|s)$$

where λ is the interpolation weight ($0 \leq \lambda \leq 1$). Optimization of λ values could be difficult in practice, since the mixture components are not anymore included in (1) (Foster and Kuhn, 2007). Furhter details about training the adaptive log-linear and mixture models are given in (Sankaran et al., 2012).

3 Experimental Results

Table 1 provides a summary of the data sets used for the training, optimization and testing.

Table 1: Data sets for source and target domains

Domain	Number of sentence pairs			Avg. sentence length	
	Train	Dev	Test	English	French
NL	**1631153**	-	-	19.4	23.8
HR	**16444**	1000	1000	20.2	20.9
IA	**21037**	1000	1000	6.7	8.5

Three baseline systems are built for performance comparison with the domain adaptation techniques. Baseline 1 is trained and optimized using data from NL domain only. Baseline 2 is trained on NL training set and optimized on the development (Dev) set from the target domains (HR or IA). For the Human Rights domain, Baseline 3 is trained on the concatenation of the source and target training sets (NL $\cup$ HR) and optimized on HR Dev set (similarly for IA domain). All systems are evaluated on the testing sets from the target domains.

Table 2 presents the BLEU score evaluation of the domain adaptation techniques for both domains compared to that of the baselines. As shown in Table 2, the log-linear approach slightly outperforms Baseline 3 and mixture model for the adaptation to HR domain. For the adaptation to IA domain however, the mixture model approach provides a significant improvement (about 3 BLEU points) over that of the log-linear approach and Baseline 3.

The best of the domain adaptation techniques are then incorporated into the *Adaptive TRANSLI* (Adaptive Translation of Legal Information) system, 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.

The human evaluation is based primarily on the amount of time required by a professional post-

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

System	HR Fr → En	IA 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**

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 and yields to a reduced translation cost.

The post-editing effort is measured according to the Human-targeted Translation Edit Rate (HTER) (Snover et al., 2006; Specia and Farzindar, 2010). The HTER is defined as the minimum number of 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.

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

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 3 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 frist system is the Adaptive TRANSLI, wich incorporates the best performing adaptive (log-linear or mixture) models according to the BLEU score evaluation (presented in Table 2). 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 a post-editing tool that measures the time and the HTER value required to post-edit each sentence.

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

Table 4 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 4, the average time per word required by all post-editors to revise the output of Adaptive TRANSLI is 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 reduced compared to NLP current system. 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. The company would therefore 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.

As shown in Table 4, the average HTER values per sentence produced by the Adaptive TRANSLI (using the mixture model approach) for the IA domain, are lower than that of the current NLP system. However, for the HR domain the HTER values produced by the Adaptive TRANSLI (using the log-linear approach) are shown to be higher than

Table 3: Statistics about documents selected for human evaluation

Doc.	#Words	#Sentences	Avg. 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 4: Average post-editing time (in seconds) and HTER values required during human evaluation

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

that of NLP current system. This is mainly caused by few 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 an additional number of edits to translate the phrases ignored by the adaptive TRANSLI, which explains the high values of HTER for the French-English translation of the HR documents. In contrast to Adaptive TRANSLI, the current system employed at NLP is trained on large and diverse corpora in addition to the corpora from the legal domain (NL), and hence it is able to translate these general expressions. However, the overall time required to post-edit a document from the HR domain translated by the Adaptive TRANSLI remains lower than that of NLP current system, which demonstrates the high level of precision provided by the adaptive system.

4 Conclusion and Future Work

This paper presents an evaluation of the domain adaptation techniques in terms of traditional automatic evaluation metrics and human evaluations in a real-world setting. Adaption of current SMT systems to new domains would reduce the translation time and efforts while maintaining or improving translation quality. NLP Technologies’ Adaptive TRANSLI system, comprising two domain adaptation techniques based on log-linear models and mixture models, has been used to adapt a general SMT system to different legal domains.

The results of automatic evaluation have shown that Adaptive TRANSLI can yield a significant increase in the BLEU score over that of the baseline. 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.

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.

References

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL.

Atefeh Farzindar. 2009. Automatic translation management system for legal texts. In *MT Summit XII: Proceedings of the twelfth Machine Translation Summit*, pages 417–424, Ottawa, Ontario, aug.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for smt. In *Proceedings of the Second Workshop on Statistical Machine Translation*. ACL.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *In Proceedings of the Association of Computational Linguistics*. ACL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *In Proceedings of the Association for Computational Linguistics*, pages 311–318. ACL.

Fatiha Sadat, Howard Johnson, Akakpo Agbago, George Foster, Joel Martin, and Aaron Tikuisis. 2005. Portage: A phrase-based machine translation system. In *In Proceedings of the ACL Worskhop on Building and Using Parallel Texts*. ACL.

Baskaran Sankaran, Majid Razmara, Atefeh Farzindar, Wael Khreich, Fred Popowich, and Anoop Sarkar. 2012. Domain adaptation techniques for machine translation and their evaluation in a real-world setting. In *Canadian Conference on AI*, volume 7310 of *Lecture Notes in Computer Science*, pages 158–169. Springer.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Lucia Specia and Atefeh Farzindar. 2010. Estimating machine translation post-editing effort with HTER. In *AMTA 2010- workshop, Bringing MT to the User: MT Research and the Translation Industry*. The Ninth Conference of the Association for Machine Translation in the Americas, November.