Quality Estimation-guided Data Selection for Domain Adaptation of SMT

Pratyush Banerjee, Raphael Rubino, Johann Roturier[1], Josef van Genabith
CNGL, School of Computing, Dublin City University, Dublin, Ireland
{pbanerjee,rrubino,josef}@computing.dcu.ie
[1] Symantec Research Labs, Dublin, Ireland
johann_roturier@symantec.com

Abstract

Supplementary data selection is a strongly motivated approach in domain adaptation of statistical machine translation systems. In this paper we report a novel approach of data selection guided by automatic quality estimation. In contrast to the conventional approach of using the entire target-domain data as reference for data selection, we restrict the reference set only to sentences poorly translated by the baseline model. Automatic quality estimation is used to identify such poorly translated sentences in the target domain. Our experiments reveal that this approach provides statistically significant improvements over the unadapted baseline and achieves comparable scores to that of conventional data selection approaches with significantly smaller amounts of selected data.

1 Introduction

The quality of translations generated by a statistical machine translation (SMT) system depends heavily on the *amount* of available parallel training data, as well as on the *domain-specificity* of the training and target datasets (Axelrod et al., 2011). Real-life translation tasks are usually domain-specific in nature and require large volumes of in-domain parallel training data. However, such domain-specific parallel training data is often sparse or completely unavailable. In such scenarios, domain adaptation techniques are necessary to effectively leverage available out-of-domain or related-domain parallel data. Supplementary data selection (Hildebrand et al., 2005; Axelrod et al., 2011) is one such popular technique which uses out-of-domain parallel data to supplement sparse in-domain data. However, combining lots of out-of-domain data with small amounts of in-domain data might negatively affect translation quality by overwhelming the in-domain characteristics. Hence *relevant data selection* is used, where only a sub-part of the out-of-domain data, *relevant* to the target domain, supplements the sparse in-domain training data.

Conventionally, the data selection process is guided by all available monolingual (or bilingual) target-domain data. Sentence pairs from out-of-domain data, which are similar (in terms of a similarity metric) to the sentences in the target-domain, are chosen for adaptation with the objective of improving translation quality of all target domain sentences. However, an unadapted baseline system may already translate some target-domain sentences well, thus limiting their scope of improvement by adaptation. In contrast, the sentences poorly translated by the baseline system might have a higher potential for improvement. Utilising this category of target-domain sentences to guide the data selection process forms the primary motivation of our approach.

In order to identify the poorly translated sentences in the target domain, we utilise quality estimation (QE) techniques which involve the process of estimating how good the translation output is, through characteristic elements extracted from the source and the target texts, and also from the SMT system involved (if accessible). These features are predictive parameters derived from the text and associated with quality scores or labels, such as au-

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 101–108.

tomatic or manual evaluation scores. When the QE task consists of predicting labels, such as *good* or *bad*, for a given translation pair, classification and/or regression techniques can be used. The classification approach leads to direct label prediction, whereas the regression approach uses an acceptance threshold set on the predicted scores. In our approach, we experiment with both methods using a manually set threshold on the reference dataset. After predicting the poor translations on the target domain, the corresponding source sentences are used to select relevant supplementary parallel training data. In order to highlight the effectiveness of our approach we compare it with a standard technique of data selection based on the entire target-domain data. The experiments reveal that our approach provides improvements comparable to that of standard data selection techniques but with significantly smaller amounts of selected supplementary data.

In this paper we apply our approach to the task of adapting an SMT system to translate user-generated content in the Symantec web forums. The major challenge in translation of forum content lies in the lack of parallel forum-style training data. Hence, we utilise in-domain parallel training data in the form of Symantec translation memories (TMs) as a part of our baseline training data. Symantec TMs comprise internal documentation on Symantec products and services, while the forums consists of user discussions pertaining to the same. Hence, despite being in the same domain the TM data is clean, professionally edited and generally conforms to controlled language guidelines, whereas the forum data is often noisy, user-generated and has a wider vocabulary and colloquialisms. This difference between the training and target datasets necessitates the use of supplementary data for adaptation, thus making this an appropriate use-case for our approach.

The rest of paper is organised as follows: Section 2 presents related work relevant to our approach. Section 3 details the QE and data selection methods. Section 4 presents the experimental setup and results followed by discussions and conclusions in Section 5 and 6, respectively.

2 Related Work

QE for SMT was first applied at the word-level (Ueffing et al., 2003) and then extended to the sentence-level (Blatz et al., 2003). More recently, several studies have focused on using human scores to evaluate the translation quality in terms of post-editing effort (Callison-Burch et al., 2012) or translation adequacy (Specia et al., 2011). The promising results obtained in QE lead to interesting applications in MT, such as sentence-selection for statistical post-editing (Rubino et al., 2012) or system combination (Okita et al., 2012). In this paper, we apply QE techniques to identify *bad* translations from the target domain to drive domain adaptation by data selection.

In order to select supplementary out-of-domain data relevant to the target domain, a variety of criteria have been explored in the MT literature, ranging from information retrieval techniques (Hildebrand et al., 2005) to perplexity on 'in-domain' datasets (Foster and Kuhn, 2007). Axelrod et al. (2011) presented a technique using the bilingual difference of cross-entropy on 'in-domain' and 'out-of-domain' language models for ranking and selection by thresholding, which outperformed the monolingual perplexity based techniques. More recently, Banerjee et al. (2012) presented a novel translation-quality evaluation (rather than prediction) based data selection technique using an incremental translation model merging approach. While all these approaches select data with respect to the entire available target domain data, our approach uses only a sub-part of the same comprising potentially poorly translated sentences. Hence any of these techniques could effectively be combined with our approach. Here, we use the bilingual cross-entropy difference based approach (Axelrod et al., 2011) in our experimental setup. To the best of our knowledge, the QE-guided data selection approach is novel and is one of the primary contributions of this paper.

3 QE-based Data Selection

This section presents the details of the three individual components involved in our approach.

3.1 Automatic Quality Estimation

To distinguish between the good and the bad translations of the target-domain (English forum data in our context), we experimented with both classification as well as regression-based QE approaches. For both sets of experiments, we extract 17 features similar to the baseline QE setup

suggested by the organisers of the WMT12 shared task (Callison-Burch et al., 2012), which were shown to perform well on a post-editing effort prediction task. In our study, we want to predict the Translation Edit Rate (TER) (Snover et al., 2006) to spot *bad* translations. Given the TER scores for a set of translations, identifying the bad translations requires a threshold value, such that all sentences having TER scores above this threshold would be labelled as *bad* translation. However, a translation with a low TER score may still be considered *bad* since TER does not incorporate the notion of semantic equivalence (Snover et al., 2006).

To set the value of this threshold, we selected two sets of 50 sentences randomly from our QE En–Fr training data such that there was an overlap of 10 sentences in each. These sentences along with their manual translations, baseline SMT generated translations and TER scores were reviewed by two evaluators who are native French speakers. The objective of the manual evaluation was to identify the TER score threshold which could reliably distinguish between *good* and *bad* translations according to human judgement. Following the manual evaluation, the TER threshold value was set to 0.42 for the current task. Depending on when this thresholding value is applied, we distinguish the two QE approaches used in our experiments.

3.1.1 Classification

For the classification-based approach, since training a classifier requires labelled training data, thresholding is applied on the training data prior to training in order to directly predict the two labels. For each source sentence s and its translation t' from the training corpus, we associate the label x corresponding to the rule (1):

$$x = \begin{cases} 0 & \text{if } f(t', t) > \delta \\ 1 & \text{else} \end{cases} \quad (1)$$

where t is a translation reference, f is the evaluation function (TER in our case) and δ is the determined threshold. On unseen data, the trained classifier is used to infer one of the two labels for the translation of each source sentence. In the current classification context, we associate the labels 0 and 1 with *bad* and *good* translations, respectively.

3.1.2 Regression

Unlike the classification model, the regression model can be trained on the training data without applying the threshold initially. Once the model has been built and is used to predict the scores for an unseen set of translations, the threshold value is applied to label the data set and identify *bad* translations. However, the regression approach requires the computation of 2 different threshold values: (i) a *reference threshold* set on the test set TER scores and (ii) a *prediction threshold* which is set on the TER predicted by the regression model. Setting the reference threshold to the manually set threshold value of 0.42 and using an unseen development set randomly selected from the training data, the *prediction threshold* is set by optimising the performance of the regression model with respect to an evaluation metric (precision, recall, accuracy, etc.). In the context of our experiments, the threshold is set by optimising the F1 score with label 0 as the true positive, thus optimising both precision and recall for the bad translations.

3.2 Data Selection

In order to perform data selection, we use an approach based on the technique presented by Axelrod et al. (2011), to rank out-of-domain sentence pairs according to their relevance to our target domain. According to this approach, each sentence-pair from the out-of-domain corpora is ranked according to the formula in (2):

$$[H_{i_{src}}(s) - H_{o_{src}}(s)] + [H_{i_{trg}}(s) - H_{o_{trg}}(s)] \quad (2)$$

where $H_{i_{src}}$ and $H_{o_{src}}$ refer to the cross entropy of the source sentence on the in-domain and out-of-domain language models (LM), respectively, while $H_{i_{trg}}$ and $H_{o_{trg}}$ refer to cross-entropy of the target sentences on similar target side LMs. In contrast to the ranking sentences using only target domain LM, this technique biases towards the sentences which are both *like* the in-domain corpus and *unlike* the average of the out-of-domain corpora. The out-of-domain LMs used in this context are built on a randomly selected sub-sample of the supplementary data having the same size and vocabulary as that of the in-domain LM (both for source and target). Eventually, the sentence-pairs are sorted by the scores and the lowest-scoring sentences are selected by using a threshold.

While the bilingual cross-entropy difference based approach forms the basis of our data selection technique, we use an important variation on it to suit our context: In contrast to biasing the scores towards all target-domain sentences, our approach requires bias towards the set of potentially poorly translated target domain sentences. To allow this shift of bias, the source in-domain LM is trained only on the subset of the target-domain sentences which are poorly translated by the baseline. The source-side out-of-domain LMs on the other hand, is trained on a concatenation of the remaining target-domain sentences (well translated by the baseline) and the out-of-domain corpora. Finally, we use perplexity (instead of cross-entropy) for ranking the out-of-domain sentences in our experiments.[1] Secondly, In order to make the in-domain and out-of-domain LMs comparable, we restrict the vocabulary and the size of the out-of-domain LM to that of the in-domain LM. Hence, the modified scoring function used for ranking sentences for our experiments is given as (3):

$$[PP_{i_{src'}}(s) \quad - \quad PP_{o_{src}+i_{src''}}(s)] + [PP_{i_{trg}}(s) - PP_{o_{trg}}(s)] \quad (3)$$

where $PP_{i_{src'}}$ indicates the perplexity on the in-domain LM trained only on the source-side of the poorly translated sentences while $PP_{o_{src}+i_{src''}}$ refers to the LM trained on the remaining target-domain data and out-of-domain data. Note that the target side of the scoring remains the same, as there is no notion of good or bad translations in the target side of the bitext data.

3.3 Data Combination

Multiple techniques exist in the SMT literature to combine out-of-domain data with in-domain data. The combination could be done using instance weighting (Jiang and Zhai, 2007), or by linearly interpolating the phrase tables (Foster and Kuhn, 2007). Considering the success of linear interpolation outperforming the other techniques (Sennrich, 2012), we choose this technique to combine the two datasets.

In order to learn the interpolation weights, LMs are constructed on the target side of the in-domain training set and the selected supplementary data. These LMs are then interpolated using expectation maximisation on the target side of the devset to learn the optimal mixture weights. These weights are subsequently used to combine the individual feature values for every phrase pair from two phrase-tables using a weighted linear interpolation scheme. For the LMs, individual models trained on the in-domain and selected out-of-domain datasets are interpolated in a similar fashion with interpolation weights set on the devset.

[1] As cross-entropy and perplexity are monotonically related, they produce the same ranking.

4 Experimental Setup

4.1 Datasets and Tools

The primary in-domain training data for our baseline systems comprises En–Fr bilingual datasets from Symantec TMs. Considering the wider vocabulary of the forum content, we use the freely available Europarl (EP) version 6 (Koehn, 2005) and News Commentary (NC)[2] datasets in combination with the Symantec TMs to create a stronger second baseline model. We then use the following two freely available parallel datasets from the web, as the supplementary resources for data selection experiments:

- OpenSubtitles2011 (OPS) Corpus[3].
- MultiUN (UN) Parallel Corpus[4]

	Data Set	Line Cnt.	En. Token	Fr. Token
Bi-text Data	Symantec TM	3,659,455	72,604,817	82,046,300
	Europarl	1,924,594	52,139,148	57,837,037
	News-Comm.	134,757	3,338,552	3,917,982
	Dev	1,692	22,661	25,840
	Test	1,032	13,160	15,164
Supp. Data	MultiUN	9,010,933	227,085,145	263,051,365
	Open-Subs.	19,835,265	154,307,759	145,769,773
Mono Data	English Forum	1,276,136	19,964,837	
	French Forum	83,575		908,106

Table 1: Number of sentences and token counts for training, development, test, supplementary data and forum data sets.

Monolingual Symantec forum posts in French along with the target side of the TM, EP and NC training data serve as baseline language modelling data. All the LMs in our experiments are linearly interpolated with the weights set by expectation maximisation on the development (dev) set. Furthermore, a sizeable amount of English forum data (Banerjee et al., 2012) is used to create the source-side target-domain LM which is used both in the

[2] http://www.statmt.org/wmt12/translation-task.html
[3] http://www.opensubtitles.org/
[4] http://www.euromatrixplus.net/multi-un/

data selection and in determining the set of potentially poorly translated sentences in the target-domain. The dev and test sets are randomly selected from this English forum data and manually translated by professional translators. Table 1 reports the statistics on all the datasets used in all our experiments.

The SMT system used in our experiments is based on the standard phrase-based SMT toolkit: Moses (Koehn et al., 2007). The feature weights are tuned using Minimum Error Rate Training (Och, 2003) on the devset. All the LMs in our experiments are created using the IRSTLM (Federico et al., 2008) language modelling toolkit. Finally, translations of the test sets in every phase of our experiments are evaluated using BLEU, METEOR (Banerjee and Lavie, 2005) and TER (Snover et al., 2006) scores.

The classification and regression models used in the QE component of our approach are based on Support Vector Machines (SVMs) (Joachims, 1999) using Radial Basis function (RBF) kernels. We use the LibSVM toolkit:[5] a free open source implementation of the technology, for all our classification/regression model training and predictions. In order to tune the features of the SVM-based classification and regression models the grid search functionality associated with LibSVM is used. The process of feature extraction is performed using an inhouse tool.

4.2 QE Results

As stated in Section 3.1, we use both the classification and regression approaches to the QE task. For both models, we use 1200 randomly selected sentences from the devset (Table 1), to actually train the model and the remaining 492 sentences to optimise the SVM parameters using grid search. The classification and regression models are both evaluated on the available testset.

For the classification-based QE, we label the training data sentences by using the manually set threshold (0.42) on their TER scores. The testset is also labelled likewise. Once the SVM parameters have been set and the model has been trained, it is used to classify the testset and the resulting predictions are compared to that of the reference predictions. For the regression setup, the model is trained using the training data and associated TER scores, and this model is used to predict the TER scores on the testset. Comparing the predicted TER scores with the true TER scores for the testset, helps us predict the performance of the regression model in terms of root-mean-squared-error (RMSE) and minimum-average-error (MAE). However once the predictions are achieved, both the predictions and the reference TERs are converted to class-label representations by applying the *prediction* and *reference thresholds*, respectively. This allows us to compare the effect of the regression approach in terms of the same metrics (F1 score) used to evaluate the classifier-based approach. For the regression setup, the prediction threshold is set by optimising the F1 score on the regression-model predictions on the devset. Figure 1 shows the variation of F1 scores for different values of the prediction threshold, and our choice of threshold value of 0.4 corresponding to the best F1 score.

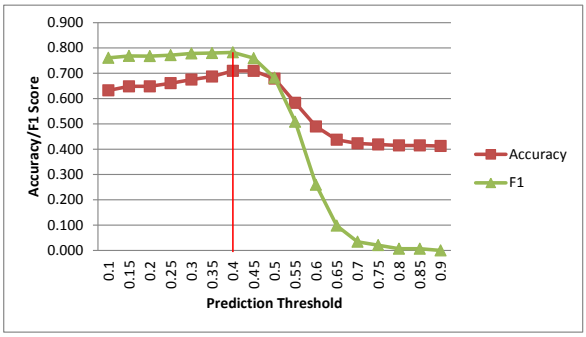

Figure 1: Variation of F1 score with prediction threshold on devset for Regression Setup.

Table 2 presents the F1 score and accuracy results on the testset for both the classification and regression setups. The accuracy and F1 scores for the regression setup correspond to an RMSE value of 0.2899 and an MAE value of 0.2104. The final column in the table indicates the percentage of the English forum data labelled as *bad* translations by the QE setup.

Configuration	Accuracy	F1 Score	% on Forum
Classification	75.2	0.8028	83.2
Regression	72.8	0.7860	83.4

Table 2: Accuracy and F1 scores on testset using classification and MAE and RMSE using regression on testset.

The results in Table 2 clearly show that using binary classification we achieve a higher accuracy on the QE task. Hence we use this particular configuration as the choice of our QE approach in order to identify potentially bad translations for data

[5]http://www.csie.ntu.edu.tw/ cjlin/libsvm/

selection. This corresponds to 83.4% (1,062,243 sentences) of the forum sentences being labelled as potentially badly translated.

4.3 SMT Results

In order to compare the effect of our approach to that of more conventional approaches, we conduct experiments on the following 5 models:

1. **BL1**: A baseline SMT model trained only on Symantec TMs.
2. **BL2**: A baseline SMT model trained on concatenated data from Symantec TMs, EP and NC parallel data sets.
3. **Full**: Using the entire supplementary datasets (either OPS or UN) in combination with the baseline (BL2).
4. **PPD**: Selecting supplementary data for the baseline (BL2) using the entire target-domain as the reference set with bilingual difference of cross-entropy (Axelrod et al., 2011).
5. **QESel**: Using our proposed approach of data selection by modified bilingual difference of perplexity (Equation 3).

We use two baseline configurations, where *BL1* is trained only on Symantec TMs while *BL2* uses additional (out-of-domain) parallel data to address data sparseness issues in the in-domain corpus. After ranking the supplementary sentence pairs using the *PPD* and *QESel* approaches, we need a threshold value to select only a section of the selected data for adaptation. In order to compare the relative effects of the two approaches on the same amount of data, we used 6 different threshold values approximately aimed at 10%, 20%, 30%, 40%, 50% and 60% of the entire datasets.

The individual translation and language models trained on these selected datasets are finally combined with the baseline models using linear interpolation techniques detailed in Section 3.3. Table 3 presents the BLEU, METEOR and TER scores for all the different configurations used in our experiments. For the *PPD* and *QESel* configurations, we present the scores for all the six sub-configurations corresponding to different sizes of the selected data. Best scores for the *QESel* and *PPD* configurations are in bold, with $*$ and † representing statistical significance over the baseline (BL2) and *Full* configurations, respectively.

The two baseline scores in Table 3 clearly indicate that *BL2* is a stronger baseline with the improvements over *BL1* being statistically significant at the p=0.05 level using bootstrap resampling (Koehn, 2004). Hence, the subsequent models are evaluated with respect to the stronger baseline (*BL2*) scores. The results show that using the supplementary data even without data selection (*Full* configuration) improves the translation quality scores. Using the UN as the supplementary data we observe a gain of 0.36 absolute BLEU points while the gain is 0.67 absolute when using OPS as the supplementary source. While the gain from using OPS as supplementary data source is statistically significant, the improvement provided by the UN datasets is not significant.

Config.		UN			OPS		
		BLEU	METEOR	TER	BLEU	METEOR	TER
BL1		31.15	48.47	0.5636	31.15	48.47	0.5636
BL2		32.27	50.19	0.5551	32.27	50.19	0.5551
Full		32.63	49.92	0.5518	*32.94	50.06	0.5460
PPD	10%	*32.75	50.03	0.5516	*32.90	50.27	0.5524
	20%	32.59	50.19	0.5518	*33.06	50.31	0.5460
	30%	*32.87	49.93	0.5473	*33.25	50.45	0.5446
	40%	*32.93	50.11	0.5489	*33.13	50.43	0.5460
	50%	*†**33.07**	50.18	0.5432	*†**33.52**	50.55	0.5450
	60%	32.59	49.93	0.5520	*33.06	50.32	0.5463
QESel	10%	*32.86	50.14	0.5458	*33.08	50.38	0.5448
	20%	*32.88	50.16	0.5487	*†**33.59**	50.96	0.5360
	30%	*†**33.19**	50.41	0.5383	*†33.46	50.63	0.5391
	40%	*†33.13	50.24	0.5451	*†33.39	50.49	0.5442
	50%	*32.84	50.21	0.5456	*†33.53	50.70	0.5451
	60%	*32.79	50.24	0.5489	*33.21	50.53	0.5448

Table 3: Testset BLEU, METEOR and TER scores for the different data selection configurations.

Using the *PPD* approach, we observe an improvement over the *BL2* baseline and the *Full* configuration using only a fraction of the datasets in most cases. For UN, using 50% of the data, we observe improvements of 0.8 and 0.44 absolute BLEU points over the baseline and *Full* configurations, respectively. The improvement figures are 1.25 and 0.58 absolute BLEU points over the baseline and Full configurations, respectively, using only 50% of the OPS dataset. All these improvements are statistically significant. METEOR and TER also follow a similar trend of improvement compared to BLEU.

The *QESel* approach, also provides statistically significant improvements over the *BL2* baseline for all sections of the full datasets. Again scores improve significantly over the *Full* configuration for most of the fraction of datasets used in our experiments. We observe an improvement of 0.92 and 0.56 absolute BLEU points using only 30% of the

UN data over the *BL2* baseline and *Full* scores, respectively. Using 20% of the entire OPS dataset, we observe improvements of 1.32 and 0.65 absolute BLEU points over the *BL2* baseline and *Full* scores, respectively. All these improvements are statistically significant at the p=0.05 level. The other evaluation metric scores also follow a similar trend of improvements. The *QESel* approach is also observed to consistently outperform the corresponding *PPD* scores for similar sizes of the supplementary datasets (the only exception being the 50% scores for UN).

5 Discussion

Comparing the improvements obtained by the two data selection approaches in Section 4.3, we observe that the *QESel* method achieves the best scores using significantly smaller amounts of data compared to the PPD approach. The *QESel* approach achieves the best improvements with only 30% and 20% of the supplementary data, for UN and OPS datasets, respectively, compared to the 50% data selected by the *PPD* approach. Furthermore, this approach provides scores which are consistently higher than the corresponding *PPD* approach for the same amount of selected sentences. Since the *QESel* approach is driven only by the poorly translated sentences in the target-domain, it prioritises the supplementary sentence pairs relevant to them. In contrast, the *PPD* approach has no particular preference towards such supplementary sentence pairs. As a consequence, selecting the top sentence pairs using the *QESel* approach improves only the previously poorly translated sentences, while *PPD* aims at uniformly improving all the target-domain sentences in general. This difference causes the *QESel* approach to achieve higher translation scores with lesser amounts of data in the current context.

To further illustrate our point, in Table 4 we present two example sentences from our testset whose *BL2* translations are labelled *good* and *bad* by the QE classifier, along with their *PPD* and *QESel* translations. The first example shows that the *PPD* approach leads to a better syntax compared to the baseline and the *QESel* approach by ordering *Pouvez-vous* properly for an interrogative sentence. Also, the verb *permettre* is in its infinitive form which is correct in this context, while the same verb is wrong in the *QESel* translations.

Good Translation	SRC	Re : Can you make it possible for users to delete their account ?
	REF	Re : Pouvez-vous accorder aux utilisateurs le droit de supprimer leur compte ?
	BL2	Re : Vous pouvez vous permettent aux utilisateurs de supprimer son compte ?
	PPD	Re : **Pouvez-vous** vous **permettre** aux utilisateurs de supprimer son compte ?
	QESel	Re : Est-ce que vous permettent aux utilisateurs de supprimer son compte ?
Bad Translation	SRC	Looks to me like the " Restart " button is highlighted - and I had just restarted (not done any update) .
	REF	Il me semble que le bouton " Redémarrer " est en surbrillance et je venais juste de redémarrer (et non d' effectuer des mises à jour) .
	BL2	On dirait que le " Redémarrer " bouton est mis en évidence - et j' ai redémarré (pas fait une mise à jour) .
	PPD	Il me semble que le " Redémarrer " bouton est mis en évidence - et j' ai redémarré (pas fait une mise à jour) .
	QESel	Il me semble que **le bouton " Redémarrer "** est mis en évidence - et j' **avais** redémarré (pas fait une mise à jour) .

Table 4: Example sentences and their translations.

This example shows how for sentences with decent baseline translation, the *PPD* approach performs better than the *QESel* approach. The second example on the other hand, shows that the *QESel* method leads to a better word ordering and keeps the correct past tense for the verb *had restarted* translated as *avais redémarré*, in comparison to *PPD* translations. The *BL2* translation for the example being *bad*, the focussed data selection by *QESel* improves it further than the conventional *PPD* approach.

Furthermore, our experiments reveal that the OPS datasets provide better improvements using both data selection methods in contrast to the UN corpus. This may be due to the informal and colloquial nature of the OPS corpus which makes it more appropriate to adapting SMT models for translating forum content.

6 Conclusion and Future Work

In this paper we presented a QE-guided data selection approach for domain adaptation of SMT systems using only the part of the target-domain data that is poorly translated by the baseline system. Our experiments revealed that this approach performs significantly better than the unadapted baseline model as well as the model using the entire supplementary data without any data selection. Furthermore, this approach also achieves similar or better improvements to that of conventional data selection approaches with considerably smaller amounts of selected data.

Despite using a set of baseline features for the QE task, our approach shows promising results thereby indicating a number of possible future directions. Extending the set of QE features to im-

prove prediction/classification performance is the primary future direction. We would also like to investigate the effect of this approach using a finer grained classification approach. Finally a deeper investigation into sophisticated data selection and ranking schemes is necessary to further exploit the effectiveness of the approach.

Acknowledgments

This work is supported by the European Commission's Seventh Framework Programme (Grant 288769) from Science Foundation Ireland (Grant 07/CE/I1142) as part of CNGL at Dublin City University, and from Research Ireland under the Enterprise Partnership Scheme (EPSPD/2011/135).

References

Axelrod, A., X. He, and J. Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on EMNLP-11*, pages 355–362, Edinburgh, United Kingdom.

Banerjee, S. and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, MI.

Banerjee, P., S. Naskar, J. Roturier, A. Way, and J. van Genabith. 2012. Translation Quality-Based Supplementary Data Selection by Incremental Update of Translation Models. In *Proceedings of COLING-2012*, pages 149–165, Mumbai, India.

Blatz, J., E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2003. Confidence Estimation for Machine Translation. In *JHU/CLSP Summer Workshop Final Report*.

Callison-Burch, C., P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51.

Federico, M., N. Bertoldi, and M. Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech 2008*, pages 1618–1621, Brisbane, Australia.

Foster, G. and R. Kuhn. 2007. Mixture-model adaptation for SMT. In *ACL 2007: Proceedings of the Second WMT*, pages 128–135, Prague, Czech Republic.

Hildebrand, A. S., M. Eck, S. Vogel, and A. Waibel. 2005. Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval. In *Proceedings of 10th EAMT Conference*, pages 119–125, Budapest, Hungary.

Jiang, J. and C. Zhai. 2007. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th Annual Meeting of ACL*, pages 264–271, Prague, Czech Republic.

Joachims, T. 1999. Making Large-Scale SVM Learning Practical. In Schölkopf, B., C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, USA.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the Interactive Poster and Demonstration Sessions, ACL 2007*, pages 177–180, Prague, Czech Republic.

Koehn, P. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the Conference on EMNLP, (EMNLP 2004)*, pages 388–395, Barcelona, Spain.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

Och, F. J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on ACL - Volume 1*, pages 160–167, Sapporo, Japan.

Okita, T., R. Rubino, and J. van Genabith. 2012. Sentence-Level Quality Estimation for MT System Combination. In *Proceedings of the ML4HMT-12 Workshop*, page 55.

Rubino, R., S. Huet, F. Lefèvre, and G. Linarès. 2012. Statistical post-editing of machine translation for domain adaptation. *Proceedings of the European Association for Machine Translation (EAMT)*, pages 221–228.

Sennrich, R. 2012. Mixture-Modeling with Unsupervised Clusters for Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 16th Annual Conference of the EAMT (EAMT-2012)*, pages 185–192, Trento, Italy.

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

Specia, L., N. Hajlaoui, C. Hallett, and W. Aziz. 2011. Predicting machine translation adequacy. *Proceedings of MT Summit XIII*, pages 19–23.

Ueffing, N., K. Macherey, and H. Ney. 2003. Confidence Measures for Statistical Machine Translation. In *Proceedings of the MT Summit IX*.