Translation Quality-Based Supplementary Data Selection by Incremental Update of Translation Models

Pratyush Banerjee[1], Sudip Kumar Naskar[1], Johann Roturier[2], Andy Way[3], Josef van Genabith[1]
(1) CNGL, School of Computing, DCU
{pbanerjee,snaskar,josef}@computing.dcu.ie
(2) Symantec Ireland, Dublin
johann_roturier@symantec.com
(3) Capita Translation and Interpreting, Delph, UK
andy.way@capita-ti.com

ABSTRACT
Supplementary data selection from out-of-domain or related-domain data is a well established technique in domain adaptation of statistical machine translation. The selection criteria for such data are mostly based on measures of similarity with available in-domain data, but not directly in terms of translation quality. In this paper, we present a technique for selecting supplementary data to improve translation performance, directly in terms of translation quality, measured by automatic evaluation metric scores. Batches of data selected from out-of-domain corpora are incrementally added to an existing baseline system and evaluated in terms of translation quality on a development set. A batch is selected only if its inclusion improves translation quality. To assist the process, we present a novel translation model merging technique that allows rapid retraining of the translation models with incremental data. When incorporated into the 'in-domain' translation models, the final cumulatively selected datasets are found to provide statistically significant improvements for a number of different supplementary datasets. Furthermore, the translation model merging technique is found to perform on a par with state-of-the-art methods of phrase-table combination.

KEYWORDS: Statistical Machine Translation, Domain Adaptation, Supplementary Data Selection, Model Merging, Incremental Update.

Proceedings of COLING 2012: Technical Papers, pages 149–166,
COLING 2012, Mumbai, December 2012.

1 Introduction

Statistical Machine Translation (SMT) has grown to be the most dominant machine translation paradigm. A prerequisite for SMT is the availability of sizeable parallel training data. The popularity of SMT has led to the free availability of a variety of parallel corpora on the web. While some such corpora comprise data from wide-coverage domains such as politics or news, others are based on much more focused and narrower domains such as medical texts or software manuals. In order to create an SMT system for a specific domain which does not have sufficient parallel training data, domain adaptation methods are necessary to best utilise supplementary parallel training data from available out-of-domain or related-domain corpora. However, the conventional wisdom of more data being better does not always hold true for domain-specific systems (Axelrod et al., 2011). Adding a lot of out-of-domain data to an in-domain SMT system tends to overwhelm the in-domain characteristics, thereby negatively affecting translation quality. Thus relevant data selection from large supplementary out-of-domain data plays an important part in domain adaptation of SMT systems.

In this paper we focus our efforts on creating an SMT system for translating user-generated forum content in Symantec web forums. Being a multinational company, Symantec supports web forums for its products and services in multiple languages with the English forum being both the oldest and (by far) the largest with considerable amounts of relevant information. Translating the forum content enables this information to be available across all languages. Moreover, these forums have also become effective sources of self-service, thus providing an alternative to traditional customer service options (Roturier and Bensadoun, 2011). However, a major challenge in building a system for forum content translation is the lack of parallel forum data for training. To overcome this challenge, we utilise 'in-domain' (but 'out-of-style') parallel training data in the form of Symantec translation memories (TMs). Symantec TMs comprise internal documentation on its products and services along with user manuals, product descriptions and some customer service communications. The forum data on the other hand, consists of posts where customers and Symantec employees discuss and solve specific problems pertaining to specific products and services. Although the TM and the forum data are in the same domain, the TM data is professionally edited and generally conforms to Symantec's controlled language guidelines. By contrast, the forum data is often noisy, not controlled by any quality guidelines and in general having a wider vocabulary and colloquialisms. This difference between the training data and target domain necessitates the use of supplementary datasets to improve translation quality.

Given the TM-based domain-specific baseline model and an additional general-domain supplementary dataset, we iteratively select batches of sentences from the supplementary dataset and add this to the in-domain translation model of the baseline system and evaluate the translation quality in terms of automatic evaluation metrics on a development set (devset). A batch is approved for addition to the baseline model only upon improvement over the baseline evaluation metric scores. In order to incrementally and rapidly retrain and evaluate the evolving translation models with each additional batch of sentences, a translation model is estimated for each batch under consideration in isolation and subsequently merged with the larger translation model using a novel phrase-table merging mechanism.

Prior to the iterative batch selection process, the supplementary training data is ranked using perplexity (normalised with sentence length) with respect to a source-side forum data language model. This technique allows the selection of batches of sentence pairs from the supplemen-

tary data with perplexity scores within a close range. Our experiments are conducted for the English–French (En–Fr) and English–German (En–De) language pairs. We use three different freely available parallel corpora as supplementary sources of data. Our experiments show that the selected supplementary datasets when incorporated into the baseline translation model consistently improve translation quality over the baseline translations, for different supplementary data sources. Comparing our data selection method with existing data selection approaches confirms the superiority of our approach in terms of translation quality improvements. In addition to the data selection, we develop a phrase table merging technique as an efficient alternative to established methods of model combination. We compare our technique of model combination to the traditional approach of static retraining, use of multiple translation models (Koehn and Schroeder, 2007) as well as mixture modelling with linear interpolation (Foster and Kuhn, 2007) to find that our technique performs at par with most of these techniques in terms of translation quality.

While the translation quality based data selection technique performs well in the experiments presented in this paper, there is a risk that the approach may overfit on the small devsets used (small devsets are a typical situation in real-life domain adaptation scenarios). In particular, this can happen if the set is not 'fully' representative of the target domain in question. Hence the evaluation during the iterative data selection phase should ideally be carried out for multiple devsets and the intersection of selected datasets from each run should be used. However generating multiple devsets for a given target domain (here user forums) is prohibitively expensive involving considerable manual effort. To alleviate this issue, the source data of the devset selected for the set of experiments reported here, is randomly chosen from a large collection of the target domain data and is ensured to be truly representative of the the target domain in terms of meta-statistics.[1] Furthermore, due care is taken during the manual translation process to preserve the characteristics of the target domain.

The rest of the paper is organized as follows: Section 2 reviews related work relevant to the task. Section 3 introduces our approach of data selection and phrase-table merging. Section 4 presents the experimental setup for our and comparative approaches. Section 5 presents the results and analysis followed by conclusions and future work.

2 Related Work

The idea of supplementary data selection from related or unrelated domains to boost the performance of sparse 'in-domain' models has been widely practised in domain adaptation of SMT (Eck et al., 2004). A wide variety of criteria for data selection has been explored ranging from information retrieval techniques (Hildebrand et al., 2005) to perplexity or cross-entropy on 'in-domain' datasets (Foster and Kuhn, 2007; Banerjee et al., 2011). Out-of-vocabulary (OOV) words with respect to training data (Daume III and Jagarlamudi, 2011; Banerjee et al., 2012) are used to mine supplementary data sources for adaptation. (Axelrod et al., 2011) presents a technique of using the difference in cross-entropy of the supplementary sentence pairs on 'in-domain' and 'out-of-domain' datasets for ranking and selection by thresholding. All these techniques rely on selecting supplementary data based on its similarity with the target domain using different measures of similarity like perplexity or OOV word rate. However, perplexity reduction often does not correlate with translation quality improvement (Axelrod, 2006). In this paper we address this issue head-on by directly using translation quality as a

[1]The parameters used are average sentence length, average type-token ratio, average stop word and function word ratio and the standard deviations of the same measures.

guide for data selection. To the best of our knowledge this is a novel approach and one of the main contributions of the paper.

In order to facilitate incremental retraining, we develop a phrase-table merging mechanism that is used to incrementally merge small phrase-tables estimated on incremental batches of supplementary dataset. Incremental updates of translation models have been attempted using a stepwise online expectation-maximization algorithm (Cappé and Moulines, 2009) for stream-based translation models (Levenberg et al., 2010) or using suffix arrays (Callison-Burch et al., 2005) to store the source–target alignments in memory. Our approach differs from these methods primarily in how we update translation model probabilities. The domain-specific aspect of our experimental setup allows us to avoid costly incremental alignment estimations. Furthermore, our approach enables merging independent translation models estimated on different domain-specific word/phrase alignments providing an alternative to other model combination techniques. While simple concatenation of in-domain and out-domain data prior to (re-) training is a commonly used (but costly) technique, multiple phrase-tables (one on each domain) can directly be combined using the decoder (Koehn and Schroeder, 2007), or interpolated using linear or log-linear weighted combination using mixture modelling (Foster and Kuhn, 2007). Our phrase-table merging technique is motivated by the linear interpolation based approach, but differs in our use of phrase-counts to merge multiple phrase-pairs.

3 Incremental Data Selection and Model Merging

This section describes in detail our data selection method and the phrase-table merging technique developed for incremental model updates.

3.1 Batching Sentence Pairs in Supplementary Data

The primary objective of our experiments is to identify the sentence pairs in the 'out-of-domain' supplementary datasets which when incorporated into the 'in-domain' model would improve translation performance. Ideally, for every sentence pair in the supplementary datasets, a new translation model needs to be retrained and its performance evaluated in terms of evaluation metrics. A sentence is suitable for selection only when its inclusion improves the translation quality of the baseline system. However, to manage the scaling issue of this approach, instead of evaluating individual sentence pairs, we group a number of them together in every iteration. In addition, updating any sizeable SMT model with a single sentence pair is unlikely to produce any measurable changes in overall translation output. The supplementary datasets are initially ranked according to their normalised perplexity with respect to a language model estimated on the English user forum dataset. In every iteration, for each batch we pick up a set of sentence pairs whose perplexity lies within a small predefined range (to be supplied by the user as input). For our experiments we use an ad-hoc value of 1 for the range although a further detailed investigation on the effect of the range size on data selection is planned for future. Since perplexity is used as a measure of 'closeness' with respect to the target domain, all pairs in the selected batch have perplexity within a small range (with a value of 1) ensuring uniform closeness of all sentences within the batch with respect to the target domain.

3.2 Selection Algorithm

To decide whether a particular batch of supplementary sentence pairs is suitable for improving translation quality, we use the process outlined in Algorithm 1. The algorithm starts with a baseline translation model BL, a baseline translation score sc_0, a perplexity range r and

Algorithm 1 Supplementary data batch selection for translation performance improvement

Require: $BL \leftarrow$ Baseline Model, $sc_0 \leftarrow$ Baseline Score, $sup \leftarrow \{pp_i, src_i, trg_i\}$, $r \leftarrow$ Perplexity Range;

1: $itn \leftarrow 1$; $step \leftarrow r$;
2: $b_{itn} \leftarrow \{\}$; $i \leftarrow 1$;
3: **while** $not(EOF(sup))$ **do**
4: **if** $pp_i \leq step$ **then**
5: $b_{itn} \leftarrow b_{itn} \cup \{src_i, trg_i\}$; $i = i + 1$;
6: **else**
7: $model_{itn} \leftarrow train\_model\{b_{itn}\} \cup BL$;
8: $sc_{itn} = evaluate\_on\_dev\{model_{itn}\}$;
9: **if** $sc_{itn} \geq sc_0$ **then**
10: $BL \leftarrow model_{itn}$; $sc_0 \leftarrow sc_{itn}$;
11: **end if**
12: $itn = itn + 1$;
13: $step \leftarrow step + r$; $b_{itn} \leftarrow \emptyset$
14: **end if**
15: **end while**

a supplementary dataset comprising source and target sentence pairs along with perplexity scores. Source and target sentence pairs are batched into a group (lines 4-6) as long as their perplexity values fall below the specified range. Once the batch is selected, a new translation model is trained on the batch and the batch model is merged with the baseline model to generate an updated model $model_{itn}$ (line 7). The updated model is then used to evaluate the devset using automatic evaluation metrics (line 8) and generate a new translation score sc_{itn}. The algorithm tests if the new score is better than the previous baseline score (line 9) and if found better updates the baseline model and score with the current model and score value in the iteration. Eventually the perplexity range is extended to the next step, and the batch is cleared for accommodating the next batch of sentences (line 13). This process runs as long as there are no more batches to process. Selected batches are accumulated to produce the final supplementary dataset used for adaptation. Since the batches are ordered according to perplexity-based similarity with respect to the target domain the algorithm makes it increasingly harder for a batch to get into the final selection as (i) later batches are less similar to the targeted domain and (ii) they need to improve on a steadily improving baseline. Therefore the algorithm implements the intuition that only those parts of generic supplementary data are selected which are good enough to generate better translation quality on the devset.

A generic SMT system is usually comprised of three different statistical components: translation model (TrM), language model (LM) and a lexical reordering model (RoM). Algorithm 1 is general enough to handle updates in all these component models. However, in this paper we only report experiments with TrM and RoM model updates and use statically trained LMs (cf. Section 3.5)

3.3 Phrase-table Merging

Ideally for every iteration step, the selected batch of supplementary sentence pairs should be combined with the 'in-domain' training data of the baseline model and a new model should be estimated. Considering the computational cost involved in full retraining, clearly this is not feasible in an iterative framework. In order to facilitate an incremental approach we develop a

set of techniques to avoid full retraining by estimating a model only on the small incremental batch and then merging the models with the existing baseline models.

Word alignment estimation is the most computationally expensive process in TrM training. Thus in order to avoid re-estimation of word-alignments in every iteration, we once and for all pre-compute the word alignments on the entire supplementary dataset and use this in every iteration. This not only reduces the estimation overhead but also addresses the issue of having poor word alignments due to small amounts of parallel data in every iteration. Word-alignments are known to benefit from domain-specific over-fitting (Gao et al., 2011) which motivated us to keep our 'in-domain' (computed on Symantec TM data) and 'out-domain' (computed on supplementary dataset) word alignments separate from each other. Hence the phrase-pairs extracted for each domain (Symantec TMs or Supplementary Datasets) are only based on domain-specific word alignments estimated from the specific corpora.

To achieve lexical table merging, the standard tables are augmented with the source and target word counts (in addition to lexical probabilities). Once new lexical tables are created on the selected batch, the baseline lexical tables are scanned for shared entries and the corresponding probabilities are updated using the formulae in (1):

$$\begin{aligned} lex_{merged}(e|f) &= lex_{bl}(e|f) \times \frac{wc_{bl}(f)}{wc_{bl}(f)+wc_{inc}(f)} + lex_{inc}(e|f) \times \frac{wc_{inc}(f)}{wc_{bl}(f)+wc_{inc}(f)} \\ lex_{merged}(f|e) &= lex_{bl}(f|e) \times \frac{wc_{bl}(e)}{wc_{bl}(e)+wc_{inc}(e)} + lex_{inc}(f|e) \times \frac{wc_{inc}(e)}{wc_{bl}(e)+wc_{inc}(e)} \end{aligned} \tag{1}$$

where $lex_{bl}, wc_{bl}, lex_{inc}$ and wc_{inc} indicate the baseline lexical probability, baseline word count, incremental lexical probability and incremental word count, respectively. e and f indicate the source and target words in this context. Entries which are not shared between the base model and the batch lexical tables are simply added to the new merged lexical table. Equation 1 approximates the lexical probabilities which would result from full retraining.

Once the lexical tables have been updated, the phrase-table estimation is completed on the batch data using the merged lexical tables. Being estimated on the merged lexical table, the inverse and direct lexical weights are already up-to-date in the new phrase-table. Hence only the remaining probabilities and counts require updates. In a similar approach to the lexical table merging, every entry in the new (incremental) batch phrase-table, is compared against the older (baseline) phrase-table and the shared phrase pairs are updated by the formulae in (2):

$$\begin{aligned} \phi_{merged}(e|f) &= \phi_{bl}(e|f) \times \frac{c_{bl}(f)}{c_{bl}(f)+c_{inc}(f)} + \phi_{inc}(e|f) \times \frac{c_{inc}(f)}{c_{bl}(f)+c_{inc}(f)} \\ \phi_{merged}(f|e) &= \phi_{bl}(f|e) \times \frac{c_{bl}(e)}{c_{bl}(e)+c_{inc}(e)} + \phi_{inc}(f|e) \times \frac{c_{inc}(e)}{c_{bl}(e)+c_{inc}(e)} \end{aligned} \tag{2}$$

where $\phi_{bl}, c_{bl}, \phi_{inc}$ and c_{inc} indicate the baseline phrase translation probability, baseline phrase count, incremental phrase translation probability and incremental phrase count, respectively. e and f indicate the source and target phrases in the context. Entries which are not shared are simply copied to the merged phrase-table. Again the updates applied to the inverse and direct translation probabilities (in equation 2) are motivated by the aim to approximate the probabilities which would ideally have been generated by full retraining.

Using these merging techniques, we are able to efficiently merge the smaller incremental models to the larger baseline models to simulate the full retraining effect. Also since the actual training only happens on the smaller batches of selected data, it is computationally much faster than full retraining at every step. Note that (1) and (2) ensure that the updated lex_{merged} and ϕ_{merged} are true probabilities such that the conditions $0 \leq lex_{merged} \leq 1$ and $0 \leq \phi_{merged} \leq 1$ hold true and both probabilities sum up to 1.

3.4 Reordering Model Merging

While the basic idea behind phrase-table merging could also be applied to the re-ordering model, we choose a simpler option for re-ordering model updates. Once a new reordering model is computed on the selected batch of supplementary data, every entry is compared to the baseline reordering table, and only new entries are added to it to generate a merged RoM. For the shared entries the reordering probabilities are retained as in the baseline model. Not only does this allow faster merging of reordering models but also ensures that for common entries 'in-domain' reordering is preferred over the 'out-of-domain' ones.

3.5 Language Models

As already stated, we use statically trained LMs for all our experiments. We use 5-gram models with modified Kneser-Ney smoothing (Kneser and Ney, 1995) and interpolated back-off. With such models adding a single n-gram into an existing model affects the probability and back-off values of all n-grams in the model. Hence incremental merging of LMs can not be achieved as easily as in the case of TrMs. Accordingly, in the current experiments we use statically estimated interpolated LMs. Three different 5-gram LMs are estimated on monolingual German and French forum data, the target side of the entire TM data and supplementary datasets, respectively. We then combine them using linear interpolation. The interpolation weights are estimated by running expectation maximization (EM) (Dempster et al., 1977) on the target side of the devset.

4 Experimental Setup

In this section, we introduce the datasets, tools and software used in our experiments. We also present the experimental setups for comparing our data selection and model merging technique with established techniques in the literature.

4.1 Datasets

The training data for our baseline systems consists of En–De and En–Fr bilingual datasets in the form of Symantec TMs. Monolingual Symantec forum posts in German and French along with the target side of the TM training data serve as language modelling data. In addition, we also have about 1.1M monolingual sentences from the English forum data which is used to create the LM with respect to which the supplementary datasets are ranked. The dev and testsets are randomly selected from this English forum dataset, ensuring that they are representative of the forum data in terms of different statistics, and manually translated by professional translators. Table 1 reports the number of sentences in the different datasets along with the average sentence length (A.S.L.) used for all our experiments.

Apart from the 'in-domain' training data, we also used the following three freely available parallel corpora as supplementary datasets for our experiments.

1. Europarl (Koehn, 2005) version 6: a parallel corpus comprising of the proceedings of the European Parliament.
2. News Commentary Corpus: released as a part of the WMT 2011 Translation Task.[2]
3. OpenSubtitles2011 Corpus:[3] a collection of documents released as part of the OPUS corpus (Tiedemann, 2009).

[2] http://www.statmt.org/wmt11/translation-task.html
[3] http://www.opensubtitles.org/

	Dataset	En–De			En–Fr		
		Sent Count	En A.S.L	De A.S.L	Sent Count	En A.S.L	Fr A.S.L
Bi-text	Symantec TM	832,723	12.86	12.99	702,267	12.42	14.86
	Dev	1,000	12.91	12.20	1,000	12.91	14.99
	Test	1,031	12.75	11.99	1,031	12.75	14.69
Supp-lement.	Europarl	1,721,980	27.48	26.11	1,809,563	27.34	30.35
	News-Comm.	135,758	24.34	24.98	115,085	24.79	29.06
	Open-Subs.	4,649,247	7.61	7.16	12,483,718	8.61	8.17
Mono-lingual	English Forum	Sent Count		1,129,749	A.S.L		12.48
	German Forum			42,521			11.78
	French Forum			41,283			14.82

Table 1: Number of sentences and A.S.L. for training, dev and testsets, and target language forum datasets.

4.2 Software and Tools

The SMT system used in our experiments is based on the standard phrase-based SMT toolkit: Moses (Koehn et al., 2007). Word alignment is performed with Giza++ (Och and Ney, 2003) using the 'grow-diag-final' heuristic. The lexical, phrase and reordering tables are built on the word alignments using the Moses training scripts. The standard training scripts are modified to augment the count information in the lexical tables. The maximum phrase-length is set to 7. The automatic metric used to evaluate translation quality in the incremental setup is BLEU (Papineni et al., 2002), although the selection algorithm is general enough to accommodate any other evaluation metric. The feature weights for the log-linear combination of the features are tuned using Minimum Error Rate Training (MERT) (Och, 2003) on the devset in terms of BLEU. For the LMs used in each of our models, we used the IRSTLM (Federico et al., 2008) language modelling toolkit for estimation as well as for the linear interpolation weight computation. In order to merge interpolated weights into a single LM, we used the weighted mixing mechanism provided by SRILM (Stolcke, 2002). Once the LMs are estimated, they are binarized using KenLM (Heafield, 2011) to ensure faster multi-threaded access during the decoding phase. Finally, translations of the testsets in every phase of our experiments are evaluated using the BLEU and TER (Snover et al., 2006) metrics.

4.3 Experiments

The primary objective of the experiments is relevant data selection from supplementary parallel training data for domain adaptation. In order to evaluate the effect of our data selection technique, we compare our method with established methods in the literature. Additionally we also compare existing mechanisms to combine the selected data with the 'in-domain' data.

4.3.1 Baseline

Prior to running the incremental data selection experiments, the baseline TrMs were estimated on the 'in-domain' (Symantec TMs) datasets. The standard Moses training scripts were modified to augment the actual word counts to the existing lexical table format. The scoring mechanism of Moses was adjusted to handle the variation in the lexical table formats. This modified version of the training scripts was then used to estimate the baseline TrM only on the Symantec TM data. Three different interpolated LMs were estimated using the technique reported in Section 4.2 each with the target side of different supplementary datasets. For experiments with a particular

supplementary dataset, we used the respective interpolated LM as the baseline for fair comparison. Therefore, the baseline for each set of experiments (for every supplementary dataset) had the same TrM but different LMs. The Giza++ alignments for each of the supplementary datasets were pre-computed and used in the iterative setup.

4.3.2 Data Selection Experiments

To evaluate the quality of our data selection approach we compare the following four data selection techniques:

1. Full: The naive approach of using the full data for adaptation.
2. PP: Data selection by ranking the supplementary data using normalised perplexity with respect to the target domain and thresholding (Foster and Kuhn, 2007).
3. PPD: Using difference in cross-entropy between in-domain and out-domain datasets to rank supplementary data followed by thresholding (Axelrod et al., 2011).
4. TQS: Translation quality-based data selection (cf. Section 3).

In order to rank the supplementary dataset sentences by normalised perplexity (PP), we used a LM trained on the English forum data as the target-domain LM. For each sentence on the source side of the supplementary dataset, its perplexity is computed on the target-domain LM. Perplexity is found to have a strong correlation with the sentence length and hence we normalize the perplexity values by sentence length. Once the perplexity values are computed, they are sued to sort the sentences thereby ensuring that the sentences which are closest to the target domain appear at the top. The data selection is performed by selecting the top N sentences from this ranked corpus. The value of N is set by the number of sentences selected using our TQS method for fair comparison.

Following the technique presented in (Axelrod et al., 2011), the difference of cross-entropy based ranking (PPD) requires an out-of-domain LM in addition to the existing in-domain LM. An out-domain LM is built on a randomly selected sub-sample of the supplementary training data having the same number of sentences and the same vocabulary as the in-domain LM. A similar set of in-domain and out-domain language models are also built on the target language side using the German and the French forum datasets for in-domain LMs and random samples from supplementary datasets as the out-of-domain LMs. Eventually each supplementary data sentence is ranked according to the difference in cross-entropy with respect to the in-domain and out-of-domain LMs summed over both the source and the target languages. Like in the case of PP, the sentences are sorted by these scores and the lowest scoring sentences are selected. However in contrast to the previous case, this ranking biases towards the sentences which are both like the in-domain sentences and unlike the average of out-of-domain sentences.

The sentences selected using our translation quality-based technique (TQS) are selected in batches using the approach described in Section 3.2. In order to speed up the translation process in the iterative framework, we utilise the multi-threaded feature of the Moses decoder. Furthermore, the merged phrase-table and the reordering models were filtered using the source side of the devset to reduce memory requirements as well as ensure faster decoding. While the other two ranking techniques require the selection of a thresholding value to select an appropriate subset of the supplementary data for adaptation, our technique is designed to automatically select a subset of the same. Therefore we use the number of sentences selected by TQS methods as the thresholding value for PP and PPD selection schemes.

4.3.3 Data Combination Experiments

Once the supplementary data is selected, this data needs to be combined with the in-domain training data for adaptation. In addition to the naive approach of concatenating the selected data to the in-domain datasets and retraining the model, we investigate three configurations of model combination based on existing methods in the SMT literature.

1. Conc: The naive approach of concatenating the selected data with the in-domain data and retraining the SMT model (Foster et al., 2010).
2. Multiple phrase-table (MPT): Creating separate phrase-tables for the in-domain and the selected data and using the multiple decoding path feature of the Moses decoder (Koehn and Schroeder, 2007).
3. Linear Interpolation (LinMix): Using a weighted linear interpolation to combine the individual phrase-tables (Foster and Kuhn, 2007).
4. PTM: Using the phrase-table merging technique reported in this paper.

In the concatenation approach (Conc), the selected supplementary data is added to the in-domain training data and a new TrM is retrained from scratch. This model is then tuned using the devset and finally tested using the testset to reveal the effect of adaptation. The Multiple phrase-table (MPT) approach requires training separate phrase-tables on the in-domain and selected data and combining them using the multiple decoding feature of the Moses decoder. The decoder uses both phrase-tables to score each of the translation options during the decoding phase. The phrase pairs which occur in both the phrase-tables are separately scored using their respective phrase-tables. In the linear interpolation approach (linmix) the two phrase-tables are combined using weights in a linear interpolation scheme. 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 EM 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 the formula in (3).

$$p_{linmix}(s|t) = \lambda p_{in}(s|t) + (1-\lambda)p_{out}(s|t) \tag{3}$$

where $p_{in}(s|t)$ and $p_{out}(s|t)$ are the feature values of individual phrase pairs from the in-domain and out-of-domain phrase-tables, respectively. λ is the tunable weight between 0 and 1.

The phrase-table merging (PTM) technique outlined in Section 3 was developed to rapidly combine incremental and baseline TrMs to aid our iterative data selection method. However, here we use it as an alternative technique to combine the in-domain and out-of-domain phrase-tables. While the basic idea behind this technique is similar to that of linear interpolation, in our technique each feature is weighted according to its frequency in the respective phrase-tables in contrast to using a global weight for every feature in LinMix. Following model combination, all the models are tuned using MERT on the devset.

5 Results and Analysis

As stated in Section 4.2, the incremental data selection process is performed by evaluating translation quality in terms of BLEU scores on the devset data. Table 2 reports the baseline scores, the best scores and the number of sentences selected during the process of incremental data selection on the devset. Alongside the number of selected sentences, the percentage figures indicate the proportion of the selected sentences with respect to the entire size of the supplementary datasets as reported in Table 1. Note that the BLEU scores reported in this table

are all non-MERT scores and the supplementary data was combined with the baseline model using the PTM method.

Lang-Pair	Model	Europarl		Open-Subtitles		News-Commentary	
		BLEU	Sent #	BLEU	Sent #	BLEU	Sent #
En–De	Baseline	22.97	663,127	22.94	1,464,798	22.91	15,473
	Best	*__24.17__	38.51%	*__24.33__	31.51%	*__23.34__	11.39%
En–Fr	Baseline	31.33	571,736	31.72	1,705,273	31.16	52,797
	Best	*__31.85__	31.60%	*__32.77__	13.66%	__31.43__	45.88%

Table 2: BLEU scores on devset using incremental TrM updates and number of sentences selected.$*$ indicates statistically significant improvement at p$\leq$0.05, best scores are in bold.

The scores in Table 2 clearly show the improvements observed on the devset for both language pairs across all supplementary datasets. While the improvements obtained using the Europarl (EP) and Open-Subtitles (OPS) corpora are statistically significant at the p=0.05 level using bootstrap resampling (Koehn, 2004) for both language pairs, the News-Commentary (NC) corpus only provides significant improvement for En–De translations. Compared to the improvements obtained on the other two sets, NC improvements are much lower, which could be attributed to the smaller size of the corpus and hence consequentially the smaller size of the selected dataset. As already stated in Section 4.3.2, the number of selected sentences as reported in Table 2 for each supplementary dataset is used as the threshold values for data selection for the PP and PPD ranking methods.

5.1 Data Selection Results

The primary objective of our approach being data selection from supplementary sources, we first report the results of our data selection methods in comparison to the other data selection techniques described in Section 4.3.2. In this phase, the selected supplementary data is concatenated with the in-domain training data to train new TrMs which are then tuned using MERT on the devset. Table 3 reports the BLEU and TER scores for the different data selection techniques in addition to our own method.

	System	Europarl		Open-Subs.		News-Comm.	
		BLEU	**TER**	**BLEU**	**TER**	**BLEU**	**TER**
En–De	Baseline	21.98	0.6436	22.56	0.6312	22.10	0.6394
	PP	$*$22.69	0.6233	$*$23.03	**0.6100**	22.24	**0.6257**
	PPD	$*$22.80	0.6211	$*$23.14	0.6127	22.34	0.6405
	Full	$*$22.58	0.6246	22.67	0.6189	22.20	0.6279
	TQS	$*$§**23.10**	**0.6190**	$*$§ **23.50**	0.6122	**22.47**	0.6292
En–Fr	Baseline	31.87	0.5603	32.52	0.5474	31.82	0.5569
	PP	$*$32.73	0.5506	$*$33.18	0.5452	32.28	**0.5435**
	PPD	$*$§33.03	0.5485	$*$33.26	**0.5371**	$*$§32.38	0.5527
	Full	32.39	0.5570	32.96	0.5498	31.59	0.5545
	TQS	$*$§†‡**33.58**	**0.5410**	$*$§**33.56**	0.5424	$*$§**32.56**	0.5503

Table 3: Testset BLEU and TER scores using data selection methods. $*, \dagger, \ddagger, \S$ indicates statistically significant improvement in BLEU over baseline, PP, PPD and Full datasets, respectively.

The scores reported in Table 3 show that adding additional supplementary data to the in-domain TrMs improve translation quality scores over the baseline in nearly all cases (quality

only deteriorates over the baseline when the Full NC data is added to the En–Fr training data). The actual data selection methods (PP, PPD and TQS) provide improvements on the baseline scores as well as on the Full scores, indicating the success of the data selection process. Comparing the translation quality scores between PP, PPD and TQS, we observe that while the PPD scores are sightly better than the PP scores, the TQS method performs best, consistently improving over the other two data selection methods in terms of BLEU scores. Using EP as the supplementary corpus the TQS method provides improvements of 1.12 absolute (5.1% relative) and 1.71 absolute (5.37% relative) BLEU points over the baseline scores for En–De and En–Fr translations, respectively. With the OPS corpus, the improvement figures are 0.94 absolute (4.17% relative) and 1.04 absolute (3.2% relative) BLEU points for En–De and En–Fr translations, respectively. For the NC corpus, the method improves the baseline scores by 0.37 absolute (1.67% relative) and 0.74 absolute (2.33% relative) BLEU points for En–De and En–Fr translation, respectively. While the EP and OPS improvements are statistically significant at $p \leq 0.05$ level for both language pairs, for NC only the En–Fr improvement is statistically significant. Although the TQS method provides better scores than the PP and PPD methods on all counts, the differences are not statistically significant in most cases, except for En–Fr improvements using the EP dataset. However, when compared to the Full scores, the TQS method provides statistically significant improvements for nearly all the cases.

5.2 Data Combination Results

The results reported in Table 3 use the Conc approach (cf. Section 4.3.3) to combine the additional data to the in-domain dataset. However, combining in-domain and out-domain datasets using this approach may not always lead to the best results as is evident from the literature (Foster and Kuhn, 2007; Banerjee et al., 2011). Hence in the second phase we compare the translation quality achieved by using the different combination methods explained in Section 4.3.3. Since the data selected by the TQS method was the best-performing dataset using the BLEU scores as per Table 3, we report the results of the different data combination experiments on this particular set only. Table 4 reports the effect of different data combination methods on translation score using data selected by the TQS method.

	System	Europarl		Open-Subs.		News-Comm.	
		BLEU	TER	BLEU	TER	BLEU	TER
En–De	Conc	23.10	0.6190	23.50	0.6122	22.47	0.6292
	MPT	23.15	**0.6134**	23.25	0.6145	21.75	0.6349
	PTM	23.17	0.6161	23.78	0.6116	22.58	0.6270
	LinMix	**23.23**	0.6161	∗†**23.80**	**0.6092**	†**22.66**	**0.6249**
En–Fr	Conc	33.58	0.5410	33.56	0.5424	32.56	0.5503
	MPT	33.31	0.5418	33.34	0.5456	32.20	0.5453
	PTM	33.30	0.5473	33.71	**0.5360**	32.66	**0.5324**
	LinMix	**33.75**	**0.5391**	†**33.84**	0.5398	†**32.79**	0.5494

Table 4: Testset BLEU and TER scores using data combination methods. $*, \dagger, \ddagger$ indicates statistically significant improvement in BLEU over Conc,MPT, and PTM methods, respectively.

The translation quality scores in Table 4 confirm our assumption that concatenation is not always the best option to combine multiple datasets. The results show weighted linear interpolation to be the best-performing system for different datasets and language pairs. However, the difference in the evaluation scores between the different combination techniques are mostly

statistically insignificant. MPT is found to work better than Conc in some of the cases (for EP datasets in En–De and En–Fr) but in most cases is poorer than all the other methods. Weighted linear interpolation is known to work well in multi-domain phrase-table combination (Banerjee et al., 2011) and our experiments confirm the observation. Interestingly, using our phrase-table merging method (PTM) for model combination seems to work reasonably well for all the different datasets and language pairs. While it does not perform the best, it certainly performs at par with the other combination techniques experimented with, the differences being statistically insignificant in all cases.

Using the MPT configuration has a major advantage over the Conc approach in keeping the in-domain and out-domain phrase-tables separate. While this can really be an effective choice in some cases, this model has larger number of parameters which are difficult to optimize using MERT (Chiang et al., 2009). The linear interpolation mechanism avoids the large parameter setting by combining features from multiple tables into a single table. However, this requires the estimation of the interpolation weights and it is not very straightforward to optimize the linear weights directly in terms of translation quality. While the LinMix method uses global weights for all phrase pairs, the PTM method uses different weights based on the frequency of occurrence in each corpora. This avoids the problem of linear interpolation weight optimization as well as the large parameter setting. In our experimental setting, this method slightly underperforms with respect to LinMix, but the difference is statistically insignificant.

5.3 Combining Data Selection and Model Combination

The results in Table 4 clearly indicate that linear interpolation of phrase-tables provides the best scores among different data combination techniques at least for the datasets under consideration. Hence in the final phase we present the results on different data selection methods using linear interpolated mixture models as the combination technique in Table 5

	System	Europarl		Open-Subs.		News-Comm.	
		BLEU	**TER**	**BLEU**	**TER**	**BLEU**	**TER**
En–De	PP	22.96	0.6212	23.13	0.6117	22.33	**0.6237**
	PPD	23.05	0.6225	23.26	0.6188	22.41	0.6258
	Full	22.73	0.6219	22.83	0.6177	22.25	0.6319
	TQS	§**23.23**	**0.6161**	†‡§**23.80**	**0.6092**	**22.66**	0.6249
En–Fr	PP	33.00	0.5476	33.25	0.5412	32.41	0.5487
	PPD	33.29	0.5429	33.32	**0.5379**	32.62	**0.5481**
	Full	32.80	0.5467	33.01	0.5518	31.96	0.5558
	TQS	†§**33.75**	**0.5391**	†‡§**33.84**	0.5398	§**32.79**	0.5494

Table 5: Testset BLEU and TER scores with LinMix as combination method. †, ‡, § indicate statistically significant BLEU improvements over PP, PPD and Full scores.

Using linear interpolation to combine the models built on different datasets results in a more-or-less uniform improvement in all translation quality scores for all datasets and language directions when compared to the results in Table 3. The data selected using the TQS method provides statistically significant improvements over the baseline scores as well as those using the Full dataset. Furthermore, the TQS scores are now significantly better than the PP and PPD scores for the En–Fr translation on both EP and the OPS datasets and for the En–De translations on the OPS dataset. However, the improvements are still not statistically significant for the other datasets and language pair combinations.

The overall results in Tables 3 and 5 strongly suggest the success of data selection as an adaptation technique. While adding supplementary training data widens the coverage of the TrMs, thus reducing the number of untranslated words in the translations, it also provides richer lexical translation probabilities for some phrases and words which although present in the baseline models were sparsely represented. Furthermore, we have empirically shown that our translation quality based data selection method consistently outperforms perplexity ranking-based data selection approaches. While the TQS method directly uses translation quality to select supplementary sentences, the PP and PPD methods rely on the perplexity or cross-entropy for the same task. Since perplexity or cross-entropy have low correlation with actual translation quality, sentences selected using such techniques are not guaranteed to improve translation quality. In contrast the TQS method only selects groups of sentences which improve translation quality, which is our overall objective. Hence, while using the PP or the PPD method all the top sentences from the supplementary data are chosen, the TQS method discards a few of the top batches as they fail to improve translation quality on the devset in the iterative framework.

Conclusion and Future Work

In this paper we have introduced a novel method for supplementary data selection for domain adaptation of SMT systems. Sentence pairs are selected incrementally in batches from the supplementary out-of-domain bitext data and added to the baseline system and evaluated in terms of BLEU scores on a devset. A batch is selected only if it results in improved BLEU scores. Once all the batches in a supplementary dataset are processed, the batches that pass the selection are combined to produce the selected parallel data for domain adaptation. The data selected using this method is found to outperform other existing data selection methods in terms of translation quality on an unseen testset and for a number of supplementary datasets. Additionally we also present a phrase-table merging technique that is developed to facilitate iterative data selection. This technique is effectively used to combine multiple phrase-tables from different domains and performs on a par with other existing techniques in the field. Our experiments also show that data selection is an effective adaptation technique for translating user-generated content using TM based training data. Moreover, the relative comparison of different model or data combination strategies reveals that concatenating supplementary data to existing in-domain data may not always yield the best results and is outperformed by a linear interpolation approach.

Extending the concept of iterative incremental training to LMs is one of the prime future directions for this work. Further investigation into methods to avoid the overfitting issue is also necessary. Finally, some analysis on the effect of batch size on translation quality in an iterative setting would also be an interesting future direction. Furthermore, the phrase-table merging technique could effectively be utilised for incremental training of TrMs.

Acknowledgments

This work is supported by Science Foundation Ireland (Grant No. 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Dublin City University. We thank the reviewers for their insightful comments.

References

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*,

EMNLP '11, pages 355–362, Edinburgh, UK.

Axelrod, A. E. (2006). Factored language models for statistical machine translation. Master's thesis, University of Edinburgh.

Banerjee, P., Naskar, S. K., Roturier, J., Way, A., and van Genabith, J. (2011). Domain Adaptation in Statistical Machine Translation of User-Forum Data using Component Level Mixture Modelling. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 285–292, Xiamen, China.

Banerjee, P., Naskar, S. K., Roturier, J., Way, A., and van Genabith, J. (2012). Domain adaptation in smt of user-generated forum content guided by oov word reduction: Normalization andor supplementary data? In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT-2012)*, pages 169–176, Trento, Italy.

Callison-Burch, C., Bannard, C., and Schroeder, J. (2005). A compact data structure for searchable translation memories. In *Proceedings of 10th Annual Conference of European Association for Machine Translation (EAMT-2005)*, pages 59–65, Budapest, Hungary.

Cappé, O. and Moulines, E. (2009). On-line expectation-maximization algorithm for latent data models. *Journal of the Royal Statistical Society Series B*, 71(3):593–613.

Chiang, D., Knight, K., and Wang, W. (2009). 11,001 new features for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 218–226, Boulder, Colorado.

Daume III, H. and Jagarlamudi, J. (2011). Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 407–412, Portland, Oregon, USA.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38.

Eck, M., Vogel, S., and Waibel, A. (2004). Language model adaptation for statistical machine translation based on information retrieval. In *Proceedings of 4th International Conference on Language Resources and Evaluation, (LREC 2004)*, pages 327–330, Lisbon, Portugal.

Federico, M., Bertoldi, N., and Cettolo, M. (2008). IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech 2008: 9th Annual Conference of the International Speech Communication Association*, pages 1618–1621, Brisbane, Australia.

Foster, G., Goutte, C., and Kuhn, R. (2010). Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 451–459, Cambridge, MA.

Foster, G. and Kuhn, R. (2007). Mixture-model adaptation for SMT. In *ACL 2007: Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135, Prague, Czech Republic.

Gao, Q., Lewis, W., Quirk, C., and Hwang, M.-Y. (2011). Incremental Training and Intentional Over-fitting of Word Alignment. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 106–113, Xiamen, China.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, UK.

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

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume I, pages 181–184, Detroit, Michigan.

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

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit X: The 10th Machine Translation Summit*, pages 79–86, Phuket, Thailand.

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

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *ACL 2007: Proceedings of the Second Workshop on Statistical Machine Translation*, pages 224–227, Prague, Czech Republic.

Levenberg, A., Callison-Burch, C., and Osborne, M. (2010). Stream-based translation models for statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 394–402, Los Angeles, CA.

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

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics, (ACL 2002)*, pages 311–318, Philadelphia, Pennsylvania.

Roturier, J. and Bensadoun, A. (2011). Evaluation of MT Systems to Translate User Generated Content. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 244–251, Xiamen,China.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, MA.

Stolcke, A. (2002). SRILM–An extensible language modeling toolkit. In *ICSLP 2002, Interspeech 2002: 7th International Conference on Spoken Language Processing*, pages 901–904, Denver, CO.

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248.