Data Selection for Compact Adapted SMT Models

Shachar Mirkin shachar.mirkin@xrce.xerox.com
Xerox Research Centre Europe, Meylan, France

Laurent Besacier laurent.besacier@imag.fr
LIG, University Of Grenoble Alps, Grenoble, France

Abstract

Data selection is a common technique for adapting statistical translation models for a specific domain, which has been shown to both improve translation quality and to reduce model size. Selection relies on some in-domain data, of the same domain of the texts expected to be translated. Selecting the sentence-pairs that are most similar to the in-domain data from a pool of parallel texts has been shown to be effective; yet, this approach holds the risk of resulting in a limited coverage, when necessary n-grams that do appear in the pool are less similar to in-domain data that is available in advance. Some methods select additional data based on the actual text that needs to be translated. While useful, this is not always a practical scenario. In this work we describe an extensive exploration of data selection techniques over Arabic to French datasets, and propose methods to address both similarity and coverage considerations while maintaining a limited model size.

1 Introduction

Data selection (DS) is a key method for domain adaptation. It is commonly used in statistical machine translation (SMT) for selecting from a *pool* of parallel sentences a subset that is more similar to the target domain, and using it to train the translation model, the language model, or both. Data selection can also be used when more compact models are needed due to memory limitations, for instance. A wide variety of selection methods have been used over the years, where the main principle is to measure the similarity of sentences from the pool to some in-domain data, either the development or the (source side of the) test set. Such similarity is often based on information theory metrics, like perplexity, applied to either side of the training data (source or target) or – as often turned out to be more effective – to both. One problem these methods suffer from is that they "overfit" the in-domain data, since sentences with unseen n-grams are less likely to be selected. Some methods address this coverage issue by expanding the selected training data to include more instances of previously unseen or infrequent n-grams. Techniques based on information retrieval (IR) have also been widely used for data selection, especially for choosing training or tuning data based on the test set, either from parallel corpora or by mining comparable corpora to find relevant parallel phrases or sentences. This produces highly adapted models for the test set, and generally reduces the number of out-of-vocabulary (OOV) words. The limitation of such methods is that they typically rely on the availability of the text for translation before the final model is produced. While sometimes possible, this is not always the case.

This work proposes an in-depth investigation of several data selection techniques for SMT adaptation. We experiment with less conventional and challenging domains for Arabic to French translation (translation of Web blogs and of dialectal conversation transcripts), and with very little available in-domain data. Our work provides insights into the usefulness of the different selection methods to each dataset.

In addition, we propose two data selection methods that aim to ease the tension between the similarity and coverage objectives, while the need for compact models that require rather aggressive data filtering, is taken into account. In the first proposed method, an IR model is added to a perplexity-based one, resulting in a model that is ready to use with or without relying on the test set. The second, denoted AVSF, is a domain-adaptation-fitted version of VSF (Lewis and Eetemadi, 2013), an algorithm which was designed to reduce model size without jeopardizing n-gram coverage. Both proposed methods are able to show competitive results in comparison to prior DS techniques. Over one dataset, their performance is better than other assessed methods; over the other – where data selection with size constraints proved more difficult – they outperform methods with comparable training data size, demonstrating a good tradeoff between translation performance and model size.

The rest of the paper is organized as follows. Section 2 presents common approaches for data-selection in SMT. Section 3 details the setting of our experiments. Experiments based on information theory metrics are presented in Section 4. Our proposed methods are described in Sections 5 (IR-based adaptation) and 6 (adapted VSF). Section 7 shows a summary of the results and outlines our main conclusions.

2 Data selection for domain adaptation

Data selection is a way to adapt the about-to-be-trained model by using only the part of the training data that is more similar to the target domain. DS is a general approach, that has been applied to other tasks other than SMT, including Chinese word segmentation and Part-of-Speech tagging (Song et al., 2012). In SMT, DS is common practice for domain adaptation, where a subset of the bilingual parallel corpus is used for training some or all the models comprising typical phrase-based SMT models (translation, reordering and language models). Apart from better adaptation, data selection has the advantage of making the training set, and in turn - the generated models, smaller. This is a factor we consider in this work, avoiding methods or parameters with which the training data is not significantly smaller than the entire set.

In this section we review the most common techniques used for selecting training data in SMT. The assumption is that we have a small in-domain corpus, denoted $\mathcal{I}$, and a large "general" bilingual corpus, $\mathcal{G}$, a pool from which we wish to select bi-sentences that will help better translate texts of the same domain as $\mathcal{I}$.

2.1 Information theory metrics

One prominent line of research is using information theory metrics to assess each one of the sentences in the pool, $\mathcal{G}$, and choose the ones that are most similar to the provided in-domain data, $\mathcal{I}$.

Perplexity (PP) and cross-entropy (CE) *Perplexity* (PP) is perhaps the best known metric for DS for SMT. The idea is to compute the perplexity of a language model (LM) built on $\mathcal{I}$, measure the perplexity of this LM over each sentence in $\mathcal{G}$, and select the m sentences with the lowest PP scores. Gao et al. (2002) and Moore and Lewis (2010) employed this metric for language model adaptation, and in (Foster et al., 2010; Yasuda et al., 2008) it was used for translation model adaptation. The same technique is sometimes (e.g. Moore and Lewis (2010)) referred to as *cross entropy* (CE). Since, by definition, the cross-entropy is simply the exponent in the perplexity score, the base of the exponent is 2 and all scores are positive, a smaller PP means a smaller CE and vice versa. In other words, selecting m sentences with the lowest PP scores or with the lowest CE scores is equivalent.

Cross-entropy difference (CED) While PP (or its equivalent CE) were used extensively before, Moore and Lewis (2010) proposed using the difference between the cross entropies scores with respect to the in-domain corpus and with respect to the general corpus. The idea is to prefer sentences that are typical to the in-domain (i.e. low CE) and untypical to the general domain (i.e. high CE). As above, the sentences with the lowest *cross entropy difference* are selected. Axelrod et al. (2011) have extended over the CED method by computing it bilingually, taking the average of the CED over the source and over the target. They showed that bilingual CED outperforms the monolingual version, as well as other related metrics. In their experiments, best results were achieved when the in-domain and the reduced-general corpora were placed in two different translation tables. This method is still considered the state of the art in this line of research.

In our experiments we assess all the above metrics, both monolingually and bilingually, over two different datasets. We validate that bilingual data selection works best for translation models, but show that target-based selection sometimes outperforms it for language model training. Further, quite intuitively, we learn that when the general corpus is rather similar to the in-domain one, perplexity outperforms cross entropy difference. In such cases, data selection serves mainly for the purpose of reducing model size rather than performance improvement. See Section 4 for further details about these results.

2.2 Retrieval-based selection

A different approach for data selection is based on Information Retrieval (IR). Typically, the data is selected based on the source side of the test set. This makes this approach appropriate only for certain scenarios, when the text to translate is known in advance, and when the translation can be delayed until a model is constructed for it. Eck et al. (2004) used this approach to adapt the LM of an SMT model; Hildebrand et al. (2005) adapted both the translation model (TM) and the LM by using the Lemur IR system[1] for searching the training set for sentences that are similar to each of the test set sentences, and training the model over the retrieved dataset. Perplexity was used in their work to determine the size of the selection. They evaluated the effect of removing from the retrieved set duplicate sentences, caused by sentence being retrieved by multiple queries, and found it was not leading to significant performance changes. Lu et al. (2007) also used Lemur to select sentences from the bilingual corpus based on the test set. Giving more weight to duplicate sentences resulted with somewhat better results in their experiments. In our experiments, we remove duplicate sentences in order to keep the model smaller. Their results, with the top-1000 sentences for each query, improved somewhat over the baseline that used the entire bilingual corpus, while reducing the phrase-table size by almost 30%. IR was also used for related domain adaptation tasks. For instance, Chen et al. (2012) used the test set to generate an adapted tuning set, and in (Afli et al., 2012, 2013) retrieval is used to expand the training data from comparable, rather than parallel, corpora.

While methods based on information-theory metrics aim to find sentences that are similar to the development set (with no real guarantee that it will be similar to the test set), IR-based approaches can achieve tighter adaptation by specifically considering the test set. Yet, as mentioned, it is not always plausible to assume that the text to translate is available in advance. When adaptation for the test set is not possible, we risk encountering at translation time unseen or infrequent n-grams that do occur in the parallel corpus but were not selected.

2.3 Corpus coverage

Gascó et al. (2012) addressed the issue of lacking coverage by expanding the training data with more sentences that contain its infrequent source n-grams, and showed how to reduce the selec-

[1] `http://www.lemurproject.org/`

tion to the test set n-grams. Lewis and Eetemadi (2013) suggested a related approach, termed Vocabulary Saturation Filter (VSF), that filters the training data to contain only a certain, limited number of each source and target n-gram. Keeping track of the n-gram occurrences, they pass over the training data and add each sentence pair to the selected set only if the pair contains n-grams that occur less than the predefined threshold. This method ensures that every possible n-gram in the pool will be represented in the selected training data. It does not, however, make use of any in-domain data as its motivation is to reduce the model size, rather than domain adaptation. In Section 6 we show a simple way to adapt this method for this task.

3 Experimental Setting

Our experiments were conducted within the setting of the TRAD project,[2] on the Arabic-French language pair. Below we describe the datasets, corpora and the setting of our experiments.

Evaluation datasets Two different datasets were evaluated, and for each a development set was provided, consisting of approximately $10,000$ source tokens (of untokenized text). This is the only in-domain data we had at our disposal. The datasets are described below, and their size details are given in Table 1, in terms of source and target tokens, as well as in terms of the number of (bi-)sentences. These datasets are different by nature; this is reflected, for example, by the length of the sentences, evident from the table.

- **BLOGS**: Web blogs.
- **EGYP**[3]: Transcription of conversations in Egyptian dialectal Arabic.

Dataset	Sentences	Tokens-Ar	Tokens-Fr
BLOGS	398	10K	15K
EGYP	941	10K	13K

Table 1: Development sets. This is the only in-domain data we have at our disposal.

Dataset	Sentences	Tokens-Ar	# of refs
BLOGS	409	10K	4
EGYP	828	10 K	4

Table 2: Test sets.

Evaluation metric We evaluate our translations with BLEU (Papineni et al., 2002) using an official script of the evaluation campaign that removes punctuations and lowercases the detokenized output.

Corpora The list of bilingual corpora we used is given below. Table 3 shows the size of each corpora. As our pool of bi-sentences we use a concatenation of these corpora according to the listed order. This corpus, of 14.5M bi-sentences, was also used to produce baseline models in our experiments.

- **NEWS**: News commentary.
- **WIT3** (Cettolo et al., 2012): Transcribed and translated TED talks.[4]

[2] http://www.trad-campaign.org/

[3] Within the TRAD project, this dataset is referred to as ‘H5’. This is the only dataset that is not publicly available.

[4] Downloaded from https://wit3.fbk.eu/mt.php?release=2012-02

- **TRANSCRIPTS** (H4): Radio and television transcripts; standard Arabic.
- **MULTI-UN**: United Nations official documents.[5]
- **OPENSUBTITLES** Tiedemann (2012): Movie subtitles.[6]

Dataset	Sentences	Tokens-Ar	Tokens-Fr
NEWS	91K	2.2M	2.4M
WIT3	87K	1.9M	2.4M
TRANSCRIPTS	21K	561K	778K
MULTI-UN	9.9M	222.4M	285.5M
OPENSUBTITLES	4.4M	27.7M	32.3M

Table 3: Bilingual corpora for training.

SMT system and preprocessing We used Moses (Koehn et al., 2007) as our phrase-based SMT system. IRSTLM (Federico et al., 2008) was used to train 5-gram language models over the target side of the (selected) bilingual corpora. Arabic tokenization was done similarly to MADA-TOKAN (Habash et al., 2009), but due to project constraints we used a re-implementation of this tokenizer. More precisely, 300M Arabic words of the MULTI-UN corpus, segmented using MADA, were used to train a tokenizer using OpenNLP.[7] To improve tokenization of punctuations, we applied the Moses tokenizer after the Arabic segmentation was applied. The translation models are trained on lowercased tokenized text and we apply a standard detokenizer prior to evaluation.

IR system For indexing and retrieval, we used Lucene,[8] with its default settings. We indexed all unigrams and bigrams of the preprocessed Arabic side of the bilingual concatenated corpus, and used the inverted index for retrieval for both datasets. The preprocessing of the indexed corpus is identical to that applied to any other source data, such as the development or test sets.

4 Perplexity-based adaptation

First, we assess domain adaptation techniques based on information theory metrics. The concatenated corpora is our pool, $\mathcal{G}$, and each small development set, $\mathcal{D}$, is used both for tuning and for representing the in-domain data ($\mathcal{I}$). Ideally, these tasks would use different sets, but we had no additional in-domain data available, and the data we did have was too small to split.

4.1 Perplexity (PP)

We train a LM using the development set, $\mathcal{D}$, and compute its perplexity for each sentence in $\mathcal{G}$. We refrain from building an incremental language model over $\mathcal{G}$ and measuring its perplexity over $\mathcal{D}$, but rather use the more intuitive parameter m to determine in advance how many bi-sentences we wish to use for generating the model.

We experimented with selection using perplexity over the source (denoted *pp-src*), over the target (*pp-tgt*), and over both (*pp-bi*), where the sum of perplexities of a bi-sentence is used. Once scored, the m bi-sentences with the lowest perplexity scores are selected. Table 4 shows an example of these experiments, when using the BLOGS dataset and selecting from the MULTI-UN corpus to generate the SMT model. As seen in the table, DS based on both source and target

[5] http://opus.lingfil.uu.se/MultiUN.php

[6] http://www.opensubtitles.org; the corpus was downloaded from http://opus.lingfil.uu.se/OpenSubtitles2012.php

[7] https://opennlp.apache.org/

[8] http://lucene.apache.org

proves most useful, and selection based on the source alone yields particularly inferior results. This outcome was consistent over additional values of m, over different pools of bilingual data and over the EGYP dataset as well.

BLOGS		
m	Selection	BLEU$_{dev}$
100K	*pp-src*	10.56
	pp-tgt	17.18
	pp-bi	**17.85**
300K	*pp-src*	16.82
	pp-tgt	18.48
	pp-bi	**21.15**

Table 4: Perplexity-based DS results, when selecting based on the source (*src*), the target (*tgt*), or both (*bi*). m denotes the number of selected sentence-pairs. Results are computed on the BLOGS development set.

4.2 Cross-entropy difference (CED)

We assessed the methods suggested by Moore and Lewis (2010) and Axelrod et al. (2011) for monolingual or bilingual CED. As above, we used the development set, $\mathcal{D}$, as our in-domain data; for the general data, we used a random sample of $\mathcal{G}$ of the same size as $\mathcal{D}$. Following the PP results, we assessed selection using CED based on target alone and based on both source and target, but not over the source alone. We further assessed the option to select data for the TM bilingually but for the LM monolingually, using only the target side of the corpus. A subset of our results over the EGYP dataset, using perplexity and CED, is shown in Table 5.

EGYP			
m	TM	LM	BLEU$_{dev}$
1M	*pp-bi*	*pp-bi*	10.60
	ced-bi	*ced-bi*	11.18
	ced-bi	*pp-tgt*	11.36
	ced-bi	*ced-tgt*	**11.75**
2M	*pp-bi*	*pp-bi*	11.18
	ced-bi	*ced-bi*	11.59
	ced-bi	*pp-tgt*	12.06
	ced-bi	*ced-tgt*	**12.25**
3M	*pp-bi*	*pp-bi*	10.85
	ced-bi	*ced-bi*	11.91
	ced-bi	*pp-tgt*	11.99
	ced-bi	*ced-tgt*	**12.03**
14.5M	-	-	10.63

Table 5: Perplexity and CED results for the EGYP dataset. Results are computed over the development set. The shaded row shows the baseline, where all bilingual data is concatenated and no selection is applied.

From Table 5 we learn that for the EGYP dataset: (i) adapting the LM based on the target only is better than bilingual adaptation, and (ii) CED outperforms PP over any selection size. In contrast, as shown in Table 6, for BLOGS, PP is the preferred choice. Another difference between the datasets is that BLOGS requires a larger amount of selected data to reach the performance of the simple baseline. This property of the datasets surfaced all along our experiments, and indicate that the BLOGS dataset is closer to the general corpus than EGYP. Different selection strategies had to be followed.

BLOGS		
m	Selection	BLEU$_{dev}$
100K	*ced-bi*	15.95
	pp-bi	**17.85**
1M	*ced-bi*	23.17
	pp-bi	**24.92**
2M	*ced-bi*	24.39
	pp-bi	**25.39**
3M	*pp-bi*	25.80
14.5	-	26.04

Table 6: Perplexity and CED results for the BLOGS dataset, over the development set. The shaded row shows the baseline, where no selection is applied.

For subsequent experiments, we chose the 2M selection size which obtained good results in our experiments while remaining reasonable in terms of model size.[9] With respect to the selection metric, we use the best-performing one for each dataset: *ced* (*ced-bi* for TM and *ced-tgt* for LM) for EGYP, and *pp-bi* for BLOGS.

5 IR over an adapted model

Data-selection methods based on information retrieval were described in Section 2. The principle is to generate a tightly adapted model for the (source side of the) test set by obtaining parallel corpora that covers its n-grams. We implement this approach, following prior work, with some modifications. Yet, realizing that in real-world scenarios, using the test set is not always feasible, our goal is to provide a model that can also support real-time response. In this section we describe a method that achieves this goal and the experiments using it over the two datasets. The idea is to enable working in both immediate and delayed modes. To that end, we create a model that is *IR-ready*, but still *IR-independent*. That is, a model that would be relatively easy and fast to update when we receive a text to translate and can afford a short delay, but that can also perform well when an immediate translation is required.

To support IR-based adaptation we create an inverted index of the entire preprocessed source side of the bilingual corpus. All n-grams are indexed, up to a maximal predefined length (2, in our experiments). Then, we train an adapted model based on CED or PP, using some in-domain data as described in Section 4. Since this model will not be used as our final one, and we are only after its tables (phrase table, language model and reordering table), it does not need to be tuned. Next, we let another part of the in-domain data play the role of the test set. We extract n-grams from its source side, optionally ordering them (see below). We keep track of the number of occurrences of each n-gram in the retrieved set that is being collected, $\mathcal{R}$. Searching up to k instances of each n-gram, we add to $\mathcal{R}$ the source sentences retrieved by searching exact matches of the n-gram, and their corresponding target sentences, and update the counts of each n-gram that appears in them. When the next n-gram is up for search, we deduct its current count in $\mathcal{R}$ from the maximum number of requested hits. The motivation is to obtain enough instances of each n-gram while keeping the retrieved dataset small for space and

[9] We further expand over this model in subsequent experiments, and we must therefore bound it, even for BLOGS.

speed considerations. Note that this does not constrain the absolute number of occurrences of each n-gram, and in general, more frequent n-grams will tend be more frequent in $\mathcal{R}$. With the motivation of keeping the model compact, and since the results of retrieval may overlap when separate queries lead to identical retrieved sentences, we perform a bilingual de-duplication. That is, we remove all identical bi-sentences from the selected parallel corpus, allowing duplicates to remain in either source or target, but not both. The de-duplicated set, $\mathcal{R}'$, is then used to construct additional translation and language models. The two types of adaptation are then combined in a single log linear setting, where each makes up one TM or LM. That is, we add an additional TM or LM model based on the IR data rather than train models with the entire selected data. Such separation was shown to be helpful by, e.g., Axelrod et al. (2011), and enables quickly generating the models (Mirkin and Cancedda, 2013). The updated configuration, with 2 TMs and 2 LMs, is then tuned using the development set $\mathcal{D}$, producing a ready-to-use model. If test-set adaptation is possible, we apply the same IR selection over the text for translation, and use the TM and the LM generated with it to substitute those created with the previously-available in-domain data. If we can afford, time-wise, to re-tune the model, that would be the preferred choice. Yet, tuning is a lengthy process, and if the models are of similar properties, tuning may be skipped, as shown in (Mirkin and Cancedda, 2013).

The above steps are summarized in Algorithm 1. As mentioned, only a small in-domain dataset was provided for us in these experiments. Due to this constraint, the same dataset was used for multiple roles: as a development (tuning) set, and as the seed dataset, guiding the PP and IR adaptations. If more in-domain data is available, these tasks should be performed using different sets, in order to avoid overfitting. In Algorithm 1, we refer to the in-domain data as $\mathcal{D}$, regardless of its role, as was actually done in our experiments.

Algorithm 1 : IR over an adapted model

```
Input: Bilingual pool, $\mathcal{G}$; in-domain data, $\mathcal{D}$ ; optionally: text to translate, $\mathcal{T}_{src}$
Output: An adapted model
  Index the source side of $\mathcal{G}$
  Generate a PP-adapted model, using $\mathcal{D}$ { // No tuning required}

  Train IR-based models for $\mathcal{D}$:
      $\mathcal{R}_{\mathcal{D}} = \{\}$
      Extract $n$-grams from $\mathcal{D}_{src}$; order them
      For each $n$-gram, $w$
        Search for its instances
        Update counts for all $n$-grams in the retrieved set, $\mathcal{R}(w)$
        $\mathcal{R}_{\mathcal{D}} = \mathcal{R}_{\mathcal{D}} \cup \mathcal{R}(w)$
      Bilingually de-duplicate $\mathcal{R}_{\mathcal{D}}$, producing ${\mathcal{R}'}_{\mathcal{D}}$
      Train TM and LM from ${\mathcal{R}'}_{\mathcal{D}}$
  Add the new models as TM & LM in a log-linear configuration with the PP-adapted ones
  Tune the combined model

  If given $\mathcal{T}_{src}$:
      Train IR models from $\mathcal{R}_{\mathcal{T}}$ or $\mathcal{R}_{\mathcal{T}} \cup \mathcal{R}_{\mathcal{D}}$ { // as done for $\mathcal{D}$}
      Replace the TM and LM in the tuned model
```

5.1 IR experiments

To run this method, like most methods using domain adaptation with IR, one must determine the values of a set of parameters. Below we list these parameters and describe the outcomes of exploring their usage over the development sets. Here, only the retrieved set is used to construct the SMT model, without the PP-adapted one, in order to let the different parameter values be better reflected in the results.

- k, the number of requested search hits: A higher value results with more training data, at the cost of increasing model size. We experimented with k values of 100, 500 and 1000, where $k = 1000$ yielded the best results.

- n, the maximal number of tokens in the queried n-gram: A search for a unigram matches all longer n-grams that contain it. However, since we limit the number of hits per query, it is not always the case in practice. Longer n-grams potentially constitute a better match to the query, and may therefore be valuable. Still, our experiments showed that using $n = 1$ is usually sufficient, possibly since the k value we used was able to obtain enough matches.

- The IR similarity metric: we used the default metric used by Lucene: a *tf-idf* weighted cosine similarity between the query and the document (a sentence, in our case).

- The order of searched n-grams: Since the retrieved set is sensitive to the order of the search, we assessed several ordering techniques:

  (i) "As-is": the original order of n-grams in the searched set.

  (ii) Decreasing frequency: searching for the more frequent n-grams first.

  (iii) Ratio of frequencies in the search set (e.g. $mathcalD$, representing in-domain data) and the entire pool. The idea is to give priority to words that are more prominent in the in-domain set in comparison to their "regular" prominence, much like CED. Since the number of occurrences in the entire corpus may be very high, we take its squared root, and add 1 to it, to avoid division by 0: $r = \frac{freq_{\mathcal{D}}(w)}{\sqrt{freq_{\mathcal{G}}(w)+1}}$, where w is the n-gram under consideration.

  In this set of experiments, ordering based on the last option showed but a slight improvement over the default, as-is, order. Hence, we did not further reorder the n-grams prior to the search.

Another outcome of exploring the IR parameters over the development sets was that the baseline, using the entire pool, beats the IR-based selection for BLOGS (26.04 vs. 24.80 with $k = 1000$), while for EGYP, a very small retrieved set obtains as good results as the entire pool (10.65 vs. a baseline of 10.63).

So far we have selected the non-IR adaptation method and the IR parameters. We now turn to assess the proposed combined model. We compare using different sets in the retrieval table – the development set, the test set, or both concatenated. In the latter case, we add the already-obtained $\mathcal{R}_D$ to $\mathcal{R}_T$ since it provides more statistics for generating the TM and LM, without requiring additional on-the-fly retrieval. Processing the larger dataset indeed takes longer, but is still relatively limited in comparison to the entire size of the data used in the model. Each experiment using the test set used the tuning of the development set configuration with the same parameters, only replacing the phrase table and LMs.

The results are shown in Tables 7 and 8. First, the tables show that for both datasets, even when using only the development set (the $\mathcal{R}_{\mathcal{D}}$ parts of the tables), the proposed IR model improves over PP- or CED-adapted models of a similar size. In the case of EGYP, it improves over any other model we have managed to generate.[10] Second, unsurprisingly, using the test set improves translation performance. The improvement is more significant for BLOGS than EGYP. Checking the reduction of OOVs when using the test sets, we learned that the test set significantly reduced the number of output sentences with OOVs for BLOGS (by 36%, for

[10] We note that all results for this dataset of other participants in the evaluation campaign were significantly lower than the ones presented here.

EGYP			
$\mathcal{R}$	k	Sentences	BLEU$_{test}$
ced	-	2M	13.18
ced	-	3M	12.95
no selection	-	14.5M	11.20
$\mathcal{R}_{\mathcal{D}}$	500	2.27M	13.15
	1000	2.49M	13.50
$\mathcal{R}_{\mathcal{T}}$	500	2.26M	13.69
	1000	2.47M	13.97
$\mathcal{R}_{\mathcal{T}} \cup \mathcal{R}_{\mathcal{D}}$	500	2.43M	13.59
	1000	2.63M	13.94

Table 7: EGYP results over the test set using our proposed IR method over a *ced*-adapted model of 2M bi-sentences. This is the model found to perform best over the development set, where LM is adapted based on the target only (see Section 4). Its result is shown in the shaded line, as well as *ced* with 3M bi-sentences and a baseline using all of $\mathcal{G}$.

BLOGS			
$\mathcal{R}$	k	Sentences	BLEU$_{test}$
pp-bi	-	2M	28.83
pp-bi	-	3M	29.11
no selection	-	14.5M	35.49
$\mathcal{R}_{\mathcal{D}}$	500	2.59M	30.01
	1000	3.06M	31.18
$\mathcal{R}_{\mathcal{T}}$	500	2.66M	32.04
	1000	3.16M	32.42
$\mathcal{R}_{\mathcal{T}} \cup \mathcal{R}_{\mathcal{D}}$	500	3.11M	32.66
	1000	3.93M	33.04

Table 8: BLOGS results over the test set for applying IR over a *pp-bi* adapted model of 2M bi-sentences. For comparison, other shaded lines show *pp-bi* with 3M sentences and the baseline using all of $\mathcal{G}$.

$k = 1000$); for EGYP the reduction was more modest (17%). As seen in our additional results, this dataset is very different from the available bilingual corpus and much of its vocabulary does not occur in the pool at all. Adding the development set to the test set was not found helpful for EGYP, but was so for BLOGS, where, generally speaking, improvement was observed with any addition of data.

6 Adapted Vocabulary Saturation Filter (AVSF)

The proposed IR method overcomes to some extent the "overfitting" problem of PP-based techniques. Still – when test-set adaptation is not possible – it considers the development set as its core source for adaptation, and is therefore prone to have limited coverage of the actually necessary n-grams.

VSF, an algorithm suggested by Lewis and Eetemadi (2013), aims to reduce the training data by including each n-grams only a certain number of times. Thus, any n-gram, up to the determined length, that appears in the training data, also appears in its compact version. We find this algorithm suitable to compensate for drawbacks of adaptation with PP and of IR without using the test set. VSF was not designed for domain adaptation and indeed, does not make use

Egyp				
	n	Sent. order	Sentences	BLEU$_{test}$
	-	-	14.5M	11.20
VSF	1	-	841K	8.76
	2	-	6.9M	-
AVSF	1	ced-bi (14.5M)	925K	10.29
	1	ced-bi (3M)	348K	10.78
	2	ced-bi (3M)	1.63M	13.44
	2	ced-bi (4M)	2.25M	13.81

Table 9: AVSF results for Egyp, when VSF is applied over the top m bi-sentences of the ced ordered training set. $m = 14.5M$ refers to the entire bilingual corpus. In all experiments, $t = 1$.

of any domain-specific data, such as the development set. We propose an adapted version of VSF, denoted *AVSF*, to be used for domain adaptation.

The order of the data provided to VSF has a direct effect on the selected training data. Lewis and Eetemadi (2013) discussed this issue, and ordered the training data by alignment score. They were not addressing domain adaptation in that work, and the adaptation for our task is simple: first we sort the training data according to a perplexity-based metric and then apply VSF over the reordered corpus. The idea is that sentences more relevant for the domain will be selected first and will be less likely to be skipped due to n-gram saturation.

Table 9 shows experiments we conducted using this algorithm, in comparison to its non-adapted version. VSF has 2 parameters: n, the maximal length of the n-grams we are trying to cover, and t, the minimum required frequency of each n-gram. Experiments over the development sets, with smaller corpora, showed that while increasing n is useful, t does not make much difference; we therefore set t to 1 in all our experiments. Applying VSF to our (arbitrarily ordered) bilingual corpus, with $n = 2$ and $t = 1$ results with a large amount of selected data, 6.9M bi-sentences. We consider this size as contradicting one of our goals in this work – to keep the model small – and we therefore do not train a model using this selection. Yet, for AVSF, where the order is based on domain adaptation techniques, we can use $n = 2$ by applying the algorithm only over the top-m sentences of the ordered training data. This results in a smaller selected set, while guaranteeing a coverage of the more relevant part of the corpus n-grams.

Blogs				
	n	Sent. order	Sentences	BLEU$_{test}$
	-	-	14.5M	35.49
VSF	1	-	841K	30.68
	2	-	6.9M	-
AVSF	1	pp-bi (14.5M)	947K	30.37
	1	pp-bi (3M)	232K	25.17
	2	pp-bi (3M)	1.77M	30.61
	2	pp-bi (4M)	2.34M	31.86

Table 10: Blogs results when AVSF is applied over a pp ordered training set.

The results show that increasing the size of the training data, within the size limitations we imposed, is helpful. Selection based on top-m is also beneficial, and provides more flexibility with respect to the VSF parameters. AVSF therefore proves very useful, and is also efficient in terms of run time, since once the corpus has been ordered according to pp scores, as done in

Dataset	Method	Parameters	Sentences	BLEU	BLEU with $\mathcal{T}_{src}$
EGYP	Baseline	-	14.5M	11.20	-
	PP	*ced-bi* (TM), *ced-tgt* (LM), 2M	2M	13.18	-
	IR	*ced* (2M) || $\text{IR}_{\mathcal{D}}$, $k = 1000$, $n = 1$	2.49M	13.50	13.97
	AVSF	$n = 2$, $t = 1$, *ced* (4M)	2.25M	13.81	-
BLOGS	Baseline	-	14.5M	35.49	-
	PP	*pp-bi*, 3M	3M	29.11	-
	IR	*pp-bi* || $\text{IR}_{\mathcal{D}}$, $k = 1000$, $n = 1$	3.06M	31.18	33.04
	AVSF	$n = 2$, $t = 1$, *pp-bi* (4M)	2.34M	31.86	-

Table 11: Summary of prominent results of each assessed method. The || symbol denotes a separation of translation or language models. The first model listed is the one that determines the reordering table. For IR we also show the results when the test set was available for use.

any case for DS using such metrics, each of the top-m bi-sentences needs to be traversed once, and go only through n-gram extraction.

7 Summary of results and conclusions

In Table 11 we summarize the prominent results obtained for each assessed method. The results show that AVSF and the proposed IR configuration outperform the state of the art selection with CED, even with smaller model sizes. AVSF seems to be doing better in that respect, while the IR model has the advantage of being able to tightly adapt to the test set, when some delay is permitted.

In conclusion, in this work we have investigated multiple data selection methods for domain adaptation in SMT. The generated model size played an important role in our research as we tried to achieve the best possible results with models that are trained on no more than 30% of the initial bilingual corpus. Considering the different characteristics of selection techniques, we proposed two separate methods that combine existing methods in order to benefit from the advantages of each one of them. Our extension to well-known IR-based adaptation proved competitive and enables supporting two modes of operation: instant and delayed translations. Our proposed adaptation of VSF to the task at hand was demonstrated to be useful in obtaining good performance through small models. An immediate extension of these results would be to apply the IR method over the AVSF adapted model.

Assessing two distinct datasets, we learned that DS methods are not always consistent in their success over different datasets. This is naturally reflected in the parameter values, but also in the method that needs to be used. In that respect, our experiments consistently showed that when the dataset is more "similar" to the pool (as reflected, e.g. in the BLEU scores), using more data is useful and PP is preferable over CED. Our experiments revealed some hints of how to anticipate that, but further research is required to be able to predict the most effective method for the domain and the range of parameters to assess, in order to reduce the search space. Achieving that can potentially cut the adaptation effort considerably.

Acknowledgements

This work was conducted within the framework of the TRAD project, funded by the French Government, Direction générale de l'armement (DGA).

References

Afli, H., Barrault, L., and Schwenk, H. (2012). Parallel texts extraction from multimodal comparable corpora. *Advances in Natural Language Processing*, 7614:40–51.

Afli, H., Barrault, L., and Schwenk, H. (2013). Multimodal comparable corpora as resources for extracting parallel data: Parallel phrases extraction. In *Proceedings of IJCNLP*.

Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of EMNLP*.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT^3: Web inventory of transcribed and translated talks. In *Proceedings of EAMT*.

Chen, J., Devlin, J., Cao, H., Prasad, R., and Natarajan, P. (2012). Automatic tune set generation for machine translation with limited in-domain data. In *Proceedings of EAMT*.

Eck, M., Vogel, S., and Waibel, A. (2004). Language model adaptation for statistical machine translation based on information retrieval. In *Proceedings of LREC*.

Federico, M., Bertoldi, N., and Cettolo, M. (2008). IRSTLM: an open source toolkit for handling large scale language models. In *Proceedings of INTERSPEECH*.

Foster, G. F., Goutte, C., and Kuhn, R. (2010). Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of EMNLP*.

Gao, J., Goodman, J., Li, M., and Lee, K.-F. (2002). Toward a unified approach to statistical language modeling for chinese. *ACM Transactions on Asian Language Information Processing (TALIP)*, 1(1):3–33.

Gascó, G., Rocha, M.-A., Sanchis-Trilles, G., Andrés-Ferrer, J., and Casacuberta, F. (2012). Does more data always yield better translations? In *Proceedings of EACL*.

Habash, N., Rambow, O., and Roth, R. (2009). MADA+TOKAN: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt*.

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 *Proceedings of EAMT*.

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 *Proceedings of ACL Demo and Poster Sessions*.

Lewis, W. and Eetemadi, S. (2013). Dramatically reducing training data size through vocabulary saturation. In *Proceedings of WMT*.

Lu, Y., Huang, J., and Liu, Q. (2007). Improving statistical machine translation performance by training data selection and optimization. In *Proceedings of EMNLP-CoNLL*.

Mirkin, S. and Cancedda, N. (2013). Assessing quick update methods of statistical translation models. In *Proceedings of IWSLT*.

Moore, R. C. and Lewis, W. (2010). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL*.

Song, Y., Klassen, P., Xia, F., and Kit, C. (2012). Entropy-based training data selection for domain adaptation. In *Proceedings of COLING (Posters)*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Proceedings of LREC*.

Yasuda, K., Zhang, R., Yamamoto, H., and Sumita, E. (2008). Method of selecting training data to build a compact and efficient translation model. In *Proceedings of IJCNLP*.