Does more data always yield better translations?

Guillem Gascó, Martha-Alicia Rocha, Germán Sanchis-Trilles,
Jesús Andrés-Ferrer and Francisco Casacuberta
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
{ggasco,mrocha,gsanchis,jandres,fcn}@dsic.upv.es

Abstract

Nowadays, there are large amounts of data available to train statistical machine translation systems. However, it is not clear whether all the training data actually help or not. A system trained on a subset of such huge bilingual corpora might outperform the use of all the bilingual data. This paper studies such issues by analysing two training data selection techniques: one based on approximating the probability of an in-domain corpus; and another based on infrequent n-gram occurrence. Experimental results not only report significant improvements over random sentence selection but also an improvement over a system trained with the whole available data. Surprisingly, the improvements are obtained with just a small fraction of the data that accounts for less than 0.5% of the sentences. Afterwards, we show that a much larger room for improvement exists, although this is done under non-realistic conditions.

1 Introduction

Globalisation and the popularisation of the Internet have lead to a rapid increase in the amount of bilingual corpora available. Entities such as the European Union, the United Nations and other multinational organisations need to translate all the documentation they generate. Such translations happen every day and provide very large multilingual corpora, which are oftentimes difficult to process and significantly increase the computational requirements needed to train statistical machine translation (SMT) systems. For instance, the corpora made available for recent machine translation evaluations are in the order of 1 billion running words (Callison-Burch et al., 2010).

However, two main problems arise when attempting to use this huge pool of sentences for training SMT systems: firstly, a large portion of this data is obtained from domains that differ from that in which the SMT system is to be used or assessed; secondly, the use of all this data for training the system increases the computational training requirements. Despite the previous remarks, the *de facto* standard consists in training SMT systems with all the available data. This is due to the widespread misconception that the more data a system is trained with, the better its performance should be. Although the previous statement is theoretically true if all the data belongs to the same domain, this is not the case in the problems tackled by most of the SMT systems. For instance, enterprises often need to build on-demand systems (Yuste et al., 2010). In this case, since we are interested in translating some specific text, it is not clear whether training a system with all data yields better performance than training it with a wisely selected subset of bilingual sentences.

The *bilingual sentence selection (BSS)* task is stated as the problem of selecting the best subset of bilingual sentences from an available *pool of sentences*, with which to train a SMT system. This paper is concerned to BSS, and mainly two ideas are developed.

On the one hand, two BSS strategies that attempt to build better translation systems are analysed. Such strategies are able to improve state-of-the-art translation quality without the very high computational resources that are required when using the complete pool of sentences. Both techniques span through two orthogonal criteria when selecting bilingual sentences from the available pool: avoiding to introduce a bias in the original data distribution, and increasing the informativeness of the corpus.

On the other hand, we prove that among all possible subsets from the sentence pool, there is at least a small one that yields large improvements (up to 10 BLEU points) with respect to a system trained with all the data. In order to retrieve such subset, we had to use an oracle that employs information extracted from the reference translations

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 152–161,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics

only for the purpose of selecting bilingual sentences. However, references are not used at any stage within the translation system for obtaining the hypotheses. Note that although we are not able to achieve such an improvement without an oracle, this result restates the BSS problem as an interesting approach not only for reducing computational effort but also for significantly boosting performance. To our knowledge, no previous work has quantified the room of improvement in which BSS techniques could incur.

In order to assess the performance of the different BSS techniques, translation results are obtained by using a standard state-of-the-art SMT system (Koehn et al., 2007). The most recent literature defines the SMT problem (Papineni et al., 1998; Och and Ney, 2002) as follows: given an input sentence $\mathbf{f}$ from a certain source language, the purpose is to find an output sentence $\hat{\mathbf{e}}$ in a certain target language such that

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \sum_{k=1}^{K} \lambda_k h_k(\mathbf{f}, \mathbf{e}) \qquad (1)$$

where $h_k(\mathbf{f}, \mathbf{e})$ is a score function representing an important feature for the translation of $\mathbf{f}$ into $\mathbf{e}$, as for example the language model of the target language, a reordering model or several translation models. λ_k are the log-linear combination weights.

The main contributions of this paper are:

- A BSS technique is analysed, which improves the results obtained with a random bilingual sentence selection strategy when the specific domain to be translated significantly differs from that of the pool of sentences.
- Another BSS technique is analysed that, using less than 0.5% of the sentences available, significantly improves over random selection, beating a system trained with all the pool of sentences.
- We prove, by means of an oracle, that a wise BSS technique can yield large improvements when compared with systems trained with all data available.

The remaining of the paper is structured as follows. Section 2 summarises the related work. Sections 3 and 4 present two BSS techniques, namely, probabilistic sampling and recovery of infrequent n-grams. In Section 5 experimental results are reported. Finally, the main results of the work and several future work directions are discussed in Section 6.

2 Related Work

Training data selection has been receiving an increasing amount of attention within the SMT community. For instance, in (Li et al., 2010; Gascó et al., 2010) several BSS techniques, similar to those analysed in this paper, have been applied for training MT systems when there are large training corpora available. However, neither such techniques have been formalised, nor its performance thoroughly analysed. A similar approach that gives weights to different subcorpora was proposed in (Matsoukas et al., 2009).

In (Lu et al., 2007), information retrieval methods are used in order to produce different submodels which are then weighted according to the sentence to be translated. In such work, authors define the baseline as the result obtained training only with the corpus that share the same domain of the test. Afterwards they claim that they are able to improve baseline translation quality by adding new sentences retrieved with their method. However, they neither compare their technique with random sentence selection, nor with a model trained with all the corpora.

Although the techniques that are applied for BSS are often very similar to those applied for active learning (AL), both problems are essentially different. Since the AL strategies assume that the pool of sentences are not translated, they are usually interested in finding the best monolingual subset of sentences to be translated by a human annotator. In contrast, in BSS, it is assumed that a fairly large amount of bilingual corpora is readily available, and the main goal consists in selecting only those sentences which will maximise system performance.

Some works have applied sentence selection in small scale AL frameworks. These works extend the training corpora at most with 5000 sentences. In (Ananthakrishnan et al., 2010), sentences are selected by means of discriminative techniques. In (Haffari et al., 2009) a technique is proposed for increasing the counts of phrases that are considered infrequent. Both works significantly differ from the current work not only on the framework, but also on the scale of the experiments, the

proposed techniques and the obtained improvements. Similar ideas applied to adaptation problems have been proposed in (Moore and Lewis, 2010; Axelrod et al., 2011).

3 Probabilistic Sampling

As discussed in Section 2, BSS has inherently attached many meaningful links with AL techniques. Selecting samples for learning our models, incurs in a well-known difficulty in AL, the so-called *sample bias* problem (Dasgupta, 2009). This problem, which is spread to the BSS case, is summarised as the distortion introduced by the active strategy into the probability distribution underlying the training corpus. This bias forces the training algorithm to learn a distorted probability model which can significantly differ from the actual one.

In order to further analyse the sampling bias problem, consider the maximum likelihood estimation (MLE) of a probability model, $p_\theta(\mathbf{e}, \mathbf{f})$ for a given corpus of N data points,$\{(\mathbf{e}_n, \mathbf{f}_n)\}$, sampled from the actual probability distribution, $\Pr(\mathbf{e}, \mathbf{f})$. Recall that $\mathbf{e}$ denotes a target sentence whereas $\mathbf{f}$ stands for its source counterpart. MLE techniques aims at minimising the Kullback-Leibler divergence between the actual unknown probability distribution and the probability model (Bishop, 2006), defined as

$$\mathrm{KL}(\Pr \mid p_\theta) = \sum_{\mathbf{e},\mathbf{f}} \Pr(\mathbf{e}, \mathbf{f}) \log \left(\frac{\Pr(\mathbf{e}, \mathbf{f})}{p_\theta(\mathbf{e}, \mathbf{f})} \right) \tag{2}$$

When minimising, Eq. (2) is simplified to

$$\hat{\theta} = \arg\max_\theta \sum_{\mathbf{e},\mathbf{f}} \Pr(\mathbf{e}, \mathbf{f}) \log(p_\theta(\mathbf{e}, \mathbf{f})) \tag{3}$$

which is approximated by a sufficiently large dataset under the commonly hold assumption that it is independently and identically distributed according to $\Pr(\mathbf{e}, \mathbf{f})$ as

$$\hat{\theta} = \arg\max_\theta \sum_{n} \log(p_\theta(\mathbf{e}_n, \mathbf{f}_n)) \tag{4}$$

Therefore, by perturbing the sample $\{(\mathbf{e}_n, \mathbf{f}_n)\}$ with an active strategy, we are, in fact, modifying the approximation to Eq.(3) and learning a different underlying probability distribution.

In this section a statistical framework is proposed to build systems with BSS while avoiding the sample bias. The proposed approach relies in conserving the probability distribution of the task domain by wisely selecting the bilingual pairs to be used from the whole pool of sentences. Hence, it is mandatory to exclude sentences from the pool that distort the actual probability. In order to approximate the probability distribution, we assume that a small but representative corpus is available from the task domain. This corpus, referred henceforth as the *in-domain corpus*, provides a way to build an initial model which approximates the actual probability of the system. The pool of sentences will be oppositely denoted as the *out-of-domain corpus*.

The actual probability of the task domain, the so called *in-domain* probability, is approximated with the following model

$$p(\mathbf{e}, \mathbf{f}, |\mathbf{e}|, |\mathbf{f}|) = p(\mathbf{e}, \mathbf{f} \mid |\mathbf{e}|, |\mathbf{f}|) \cdot p(|\mathbf{e}|, |\mathbf{f}|) \tag{5}$$

where $p(|\mathbf{e}|, |\mathbf{f}|)$ denotes the in-domain length probability, and $p(\mathbf{e}, \mathbf{f} \mid |\mathbf{e}|, |\mathbf{f}|)$ the in-domain bilingual probability.

The length probability is estimated by MLE

$$p(|\mathbf{e}|, |\mathbf{f}|) = \frac{N(|\mathbf{e}| + |\mathbf{f}|)}{N} \tag{6}$$

where $N(|\mathbf{e}|+|\mathbf{f}|)$ is the number of bilingual pairs in the in-domain corpus such that their lengths sum up to $|\mathbf{e}|+|\mathbf{f}|$ and N denotes the total number of sentences. Note that no distinction is made between source and target lengths since the model is intended for sampling.

The complexity of the in-domain bilingual probability distribution, $p(\mathbf{e}, \mathbf{f} \mid |\mathbf{e}|, |\mathbf{f}|)$, requires a more sophisticated approximation

$$p(\mathbf{e}, \mathbf{f}/|\mathbf{e}|, |\mathbf{f}|) = \frac{\exp(\sum_k \gamma_k f_k(\mathbf{e}, \mathbf{f}))}{\mathcal{Z}} \tag{7}$$

being $\mathcal{Z}$ a normalisation constant; and where $f_k(\ldots)$ and γ_k are the features of the model and their respective parametric weights. Specifically, four logarithmic features were considered for this sampling technique: a direct and an inverse IBM model 4 (Brown et al., 1994); and both, source and target, 5-gram language models. All feature models are estimated in the in-domain corpus with standard techniques (Brown et al., 1994; Stolcke, 2002). As a first approach, the parameters of the log-linear model in Eq. (7), γ_k, were uniformly fixed to 1.

Once we have an appropriate model for the in-domain probability distribution, the proposed method randomly samples a given number of bilingual pairs from the out-of-domain corpora (the pool of sentences). The process of extending the in-domain corpus with additional bilingual pairs from the out-of-domain corpus is summarised as follows:

- Decide according to the in-domain length probability in Eq. (6), how many samples should be drawn for each length, i.e. divide the number of sentences to add into length dependent buckets.
- Randomly draw the number of samples specified in each bucket according to the in-domain bilingual probability in Eq. (7) among all the bilingual sentences that share the current bucket length.

Although the pool of sentences is typically large, it is not large enough to gather a significant amount of probability mass. Consequently, a small set of sentences accumulate most of the probability mass and tend to be selected multiple times. To avoid this awkward and undesired behaviour, the sampling is performed *without replacement*.

4 Infrequent n-gram Recovery

Another criterion when confronting the BSS task is to increase the informativeness of the training set. Thus, it seems important to choose sentences that provide information not seen in the training corpus. Note that this criterion is sometimes opposed to the one presented in Section 3.

The performance of phrase-based machine translation systems strongly relies in the quality of the phrases extracted from the training samples. In most of the cases, the inference of such phrases or rules is based on word alignments, which cannot be computed accurately when appearing rarely in the training corpus. The extreme case are the out-of-vocabulary words: words that do not appear in the training set, cannot be translated. Moreover, this problem can be extended to sequences of words (n-grams). Consider a 2-gram $f_i f_j$ appearing few or no times in the training set. Although f_i and f_j may appear separately in the training set, the system might not be able to infer the translation of the 2-gram $f_i f_j$, which may be different from the concatenation of the translations of both words separately.

When selecting sentences from the pool it is important to choose sentences that contain n-grams that have never been seen (or have been seen just a few times) in the training set. Such n-grams will be henceforth referred to as *infrequent n-grams* . An n-gram is considered infrequent when it appears less times than an infrequent threshold t. If the source language sentences to be translated are known beforehand, the set of infrequent n-grams can be reduced to those present in such sentences. Then, the technique consists in selecting from the pool those sentences which contain infrequent n-grams present in the source sentences to be translated.

Sentences in the pool are sorted by their infrequency score in order to select first the most informative. Let $\mathcal{X}$ the set of n-grams that appear in the sentences to be translated and $\mathbf{w}$ one of them; $C(\mathbf{w})$ the counts of $\mathbf{w}$ in the source language training set; and $N(\mathbf{w})$ the counts of $\mathbf{w}$ in the source sentence $\mathbf{f}$ to be scored. The infrequency score of $\mathbf{f}$ is:

$$i(\mathbf{f}) = \sum_{\mathbf{w} \in \mathcal{X}} \min(1, N(\mathbf{w})) \max(0, t - C(\mathbf{w})) \quad (8)$$

In order to avoid giving a high score to noisy sentences with a lot of occurrences of the same infrequent n-gram, only one occurrence of each n-gram is taken into account to compute the score. In addition, the score gives more importance to the n-grams with lowest counts in the training set. Although it could be possible to select the highest scored sentences, we updated the scores each time a sentence is selected. This decision was taken to avoid the selection of too many sentences with the same infrequent n-gram. First, sentences in the pool are scored using Equation (8). Then, in each iteration, the sentence $\mathbf{f}^*$ with the highest score is selected, added to the training set and removed from the pool. In addition, the counts of the n-grams present in $\mathbf{f}^*$ are updated and, hence, the scores of the rest of the sentences in the pool. Since rescoring the whole pool would incur in a very high computational cost, a suboptimal search strategy was followed, in which the search was constrained to a given set of highest scoring sentences. Here it was set to one million.

	$t=1$		$t=10$		$t=25$	
	tr	all	tr	all	tr	all
1-gr	11.6	1.3	40.5	3.5	59.9	5.1
2-gr	38	9.8	73.2	21.3	84.9	27.9
3-gr	66.8	33.5	91.1	55.7	96.4	64.9
4-gr	87.1	65.8	98.2	85.5	99.4	90.7

Table 1: Percentage of infrequent n-grams in the TED test set when considering only the TED training set (`tr`), and when adding the out-of-domain pool (`all`), for different infrequency thresholds t.

Table 1 shows the percentage of source language infrequent n-grams for the test of a relatively small corpus such as the TED corpus (for details see Section 5) when considering just the in-domain training set ($\approx 40K$ sentences) and the same percentage when adding the larger out of domain corpora. The percentages in the table have been computed separately for different values of the threshold t and for n-grams of order from 1 to 4. Note that the reduction in the number of infrequent n-grams is very high for the 1-grams but decreases progressively when considering n-grams of higher order. This indicates that the infrequent n-grams recovery technique should be very effective for lower order n-grams, but might have less effect for higher order n-grams. Therefore, and in order to lower the computational cost involved, the experiments carried out for this paper were performed considering only infrequent 1-grams, 2-grams and 3-grams.

5 Experiments

In the present Section, we first describe the experimental framework employed to assess the performance of the BSS techniques described. Then, results for the probabilistic sentence selection strategy are shown, followed by results obtained with the infrequent n-grams technique. Some example translations are shown and, finally, we also report experiments using the infrequent n-grams technique in Oracle mode, in order to establish the potential improvement for such technique and for BSS in general.

5.1 Experimental Setup

All experiments were carried out using the open-source SMT toolkit Moses (Koehn et al., 2007), in its standard non-monotonic configuration. The phrase tables were generated by means of symmetrised word alignments obtained with GIZA++ (Och and Ney, 2003). The language model used was a 5-gram with modified Kneser-Ney smoothing (Kneser and Ney, 1995), built with SRILM toolkit (Stolcke, 2002). The log-linear combination weights in Eq. (1) were optimised using Minimum Error Rate Training (Och and Ney, 2002) on the corresponding development sets.

Subset	Language	$\|S\|$	$\|W\|$	$\|V\|$
train	English	47.5K	747K	24.6K
	French		793K	31.7K
dev	English	571	9.2K	1.9K
	French		10.3K	2.2K
test	English	641	12.6K	2.4K
	French		12.8K	2.7K

Table 2: TED corpus main figures. K denotes thousands of elements. $|S|$ stands for number of sentences, $|W|$ for number of running words, and $|V|$ for vocabulary size.

Subset	Language	$\|S\|$	$\|W\|$	$\|V\|$
train	English	77.2K	1.71M	29.9K
	French		1.99M	48K
dev 08	English	2.1K	49.8K	8.7K
	French		55.4K	7.7K
test 09	English	2.5K	65.6K	8.9K
	French		72.5K	10.6K
test 10	English	2.5K	62K	8.9K
	French		70.5K	10.3K

Table 3: News Commentary corpus main figures.

Experiments were carried out on two corpora: TED (Paul et al., 2010) and News Commentary (NC) (Callison-Burch et al., 2010). TED is an English-French corpus composed of subtitles for a collection of public speeches on a variety of topics. The same partitions as in the IWSLT2010 evaluation task (Paul et al., 2010) have been used. Subtitles have been concatenated into complete sentences. NC is a slightly larger English-French corpus in the news domain. Main figures of both corpora are shown in Tables 2 and 3. As for the pool of sentences, three large corpora have been used: Europarl (Euro), United Nations (UN) and Gigaword (Giga), in the partition established for the 2010 workshop on SMT of the ACL (Callison-Burch et al., 2010). Sentences of length greater than 50 have been pruned. Table 4 shows the main figures of the tokenised and lowercased corpora.

When translating between some language pairs, there are words that remain invariable, like for example numbers or punctuation marks in the case of European languages. In fact, an easy and

Corpus	Language	$\|S\|$	$\|W\|$	$\|V\|$
Euro	English	1.25M	25.6M	81K
	French		28.2M	101K
UN	English	5M	94.4M	302K
	French		107M	283K
Giga	English	15.5M	303M	1.6M
	French		361M	1.6M

Table 4: Figures of the corpora used as sentence pool. M stands for millions of elements.

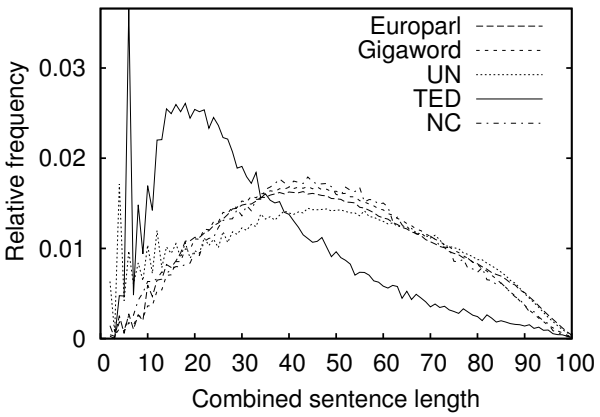

Figure 2: Combined length relative frequency.

effective technique that is commonly used is to reproduce out-of-vocabulary words from the source sentence in the target hypothesis. However, invariable n-grams are usually infrequent as well, which implies that the infrequent n-grams technique would select sentences containing such n-grams, even though they do not provide further information. As a first approach, we exclude n-grams without any letter.

Baseline experiments have been carried out for TED and NC corpora using the corresponding training set. For comparison purposes, we also included results for a purely random sentence selection without replacement. In the plots, each point corresponding to random selection represent the average of 10 repetitions. Experiments using all data are also reported, although a 64GB machine was necessary, even with binarized phrase and distortion tables.

Experiments were conducted by selecting a fixed amount of sentences according to each one of the techniques described above. Then, these sentences were included into the training data and subsequent SMT systems were built for translating the test set.

Results are shown in terms of BLEU (Papineni et al., 2001), which is an accuracy metric that measures n-gram precision, with a penalty for sentences that are too short. Although it could be argued that improvements obtained might be due to a side effect of the brevity penalty, this was not found to be true: the BSS techniques (including random) and considering all data yielded very similar brevity penalties (± 0.005), within each corpus. In addition, TER scores (Snover et al., 2006) were also computed, but are omitted for clarity purposes and since they were found to be coherent with BLEU. TER is an error metric that computes the minimum number of edits required to modify the system hypotheses so that they match the references translations.

5.2 Results for Probabilistic Sampling

In addition to the probabilistic sampling technique proposed in Section 3, we also analysed the effect of sampling only according to the combined source-reference length, with the purpose of establishing whether potential improvements were only due to the length component, or rather to the complete sampling model. Results for the 2009 test set are shown in Figure 1. Several things should be noted:

- Performing sentence selection only according to sentence lengths does not achieve better performance than random selection.
- Selecting sentences according to probabilistic sampling is able to improve random selection in the case of the TED corpus, but is not able to do so in the case of the NC corpus. Significance tests for the 500K case reported that the differences were significant in the case of the TED corpus, but not in the case of the NC corpus.
- In the case of the TED corpus, the performance achieved with the system built by sampling 500K sentences is only 0.5 BLEU points below the performance achieved by the system built with all the data available.

The explanation to the fact that probabilistic sampling is able to improve over random sampling only in the case of the TED corpus, but not in the case of NC, relies in the nature of the corpora. Although both of them belong to a very generic domain, their characteristics are very different. In fact, the NC data is very similar to the sentences in the pool, but, in contrast, the sentences present in the TED corpus have a much more different structure. This difference is illustrated in Figure 2, where the relative frequency of

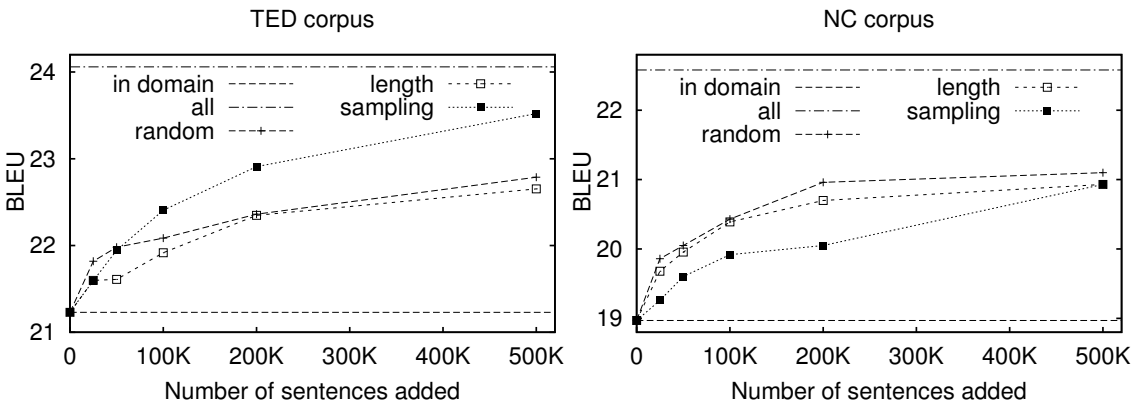

Figure 1: Effect of adding sentences over the BLEU score using the probabilistic sampling, length sampling and random selection techniques for the two corpora, TED and News Commentary. Horizontal lines represent the scores when using just the in domain training set and all the data available.

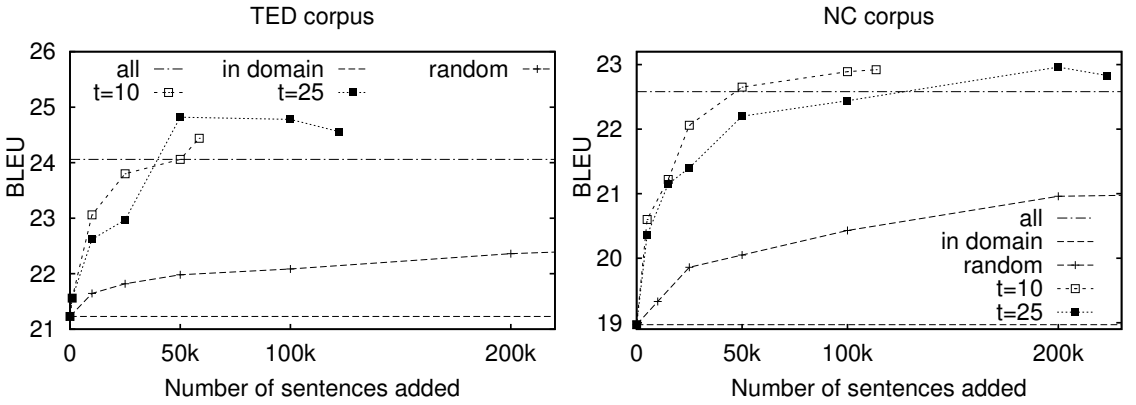

Figure 3: Effect of adding sentences over the BLEU score using the infrequent n-grams (with different thresholds) and random selection techniques for the two corpora, TED and News Commentary. Horizontal lines represent the scores when using just the in domain training set and all the data available.

each combined sentence length is shown. In this plot, it stands out clearly that the TED corpus has a very different length distribution than the other four corpora considered, whereas the NC corpus presents a very similar distribution. This implies that, when considering TED, an intelligent data selection strategy will have better chances to improve random selection than in the case of NC.

5.3 Results for Infrequent n-grams Recovery

Figure 3 shows the effect of adding sentences using the infrequent n-grams and the random selection techniques on the 2009 test set. Once all the infrequent n-grams have been covered t times, the infrequency score for all the sentences remaining in the pool is 0, and none of them can be selected. Hence, the number of sentences that can be selected for each t is limited. Although for clarity we only show results for $t = \{10, 25\}$, experiments have also been carried out for $t = \{1, 5, 10, 25\}$. Such results presented similar curves, although less sentences can be selected and hence improvements obtained are slightly lower. Several conclusions can be drawn:

- The translation quality provided by the infrequent n-grams technique is significantly better than the results achieved with random selection, comparing similar amount of sentences. Specifically, the improvements obtained are in the range of 3 BLEU points.
- Results for the TED corpus are more irregular. The best performance is achieved for $t = 25$ and 50K sentences added. In NC, the best result is for $t = 10$ and 112K.
- Selecting sentences with the infrequent n-grams technique provides better results than including all the available data. While using less than 0.5% of the data, improvements between 0.5 and 1 BLEU points are achieved.

When looking at Figure 3, one might suspect that t needs to be set specifically for a given test

set, and that results from one set are not to be extrapolated to other test sets. For this reason, we selected the best configuration in Figure 3 and used it to build a new system for translating the unseen NC 2010 test set. Such experiment, with $t = 10$ and including all sentences with score greater than 0 ($\approx$ 110K), is shown in Table 5 and evidences that improvements are actually coherent among different test sets.

technique	BLEU	TER	#phrases
in-domain	19.0	65.2	5.1M
all data	22.7	60.8	1236M
infreq. $t = 10$	23.6	59.2	16.5M

Table 5: Effect of the infrequent n-gram recovery technique for an unseen test set, when setting $t = 10$ and number of phrases (parameters) of the models.

5.4 Oracle Results

In order to analyse the potential of BSS techniques, the infrequent n-grams recovery technique in Section 4 was implemented in oracle mode. In this way, sentences from the pool were selected according to the infrequent n-grams present in the *reference* translations of the test set. Note that test references were not included into the training data as such, but were rather used to establish which bilingual sentences within the pool were best suitable for training the SMT system. In this way, we were able to establish the potential for improvement of a BSS technique. Interestingly, the SMT system trained in this way achieved 31 BLEU points on the News Commentary 2009 test set, i.e. an 8 BLEU points improvement over the system trained with all the data available. This result would have beaten all the systems that took part in the 2009 Workshop on Machine translation (Callison-Burch et al., 2009). This result is really important: although we are aware that the sentences were selected in a non-realistic manner, it proves that an appropriate BSS technique would be able to boost SMT performance in a very significant manner. Similar results were obtained with the TED and NC 2010 test sets, with 10 and 7 points improvement, respectively.

5.5 Example Translations

Example translations are shown in Figure 4. In the first example, the baseline system is not able to translate *criticised*, which is considered out-of-vocabulary. Even though random selection is able to solve this problem (luckily), it does not achieve to translate it correctly, introducing a concordance error. A similar thing happens when using probabilistic sampling, where a grammatical error is also present, and only `Infr` and `All` are able to present a correct translation. This is not only casual, since, by ensuring that a given n-gram appears at least a certain number of times t, the odds of including all possible translations of *criticised* are incremented significantly. Note that, even if the `Infr` translation is different from the reference, it is equally correct. In the second example, the baseline translation is pretty much correct, but has a different meaning (something like "and one has music"). Similarly, when including all data the translation obtained by the system means "and some music". In this case, both random and probabilistic selection present grammatically incorrect sentences, and only `Infr` is able to provide a correct translation, although pretty literal and different from the reference.

Src	the budget has also been criticised by klaus .
Bsl	le budget a également été criticised par m. klaus .
Rdm	le budget a également été critiquées par m. klaus .
PS	le budget a également été critiquée par klaus .
All	le budget a également été critiqué par klaus .
Infr	le budget a également été critiqué par klaus .
Ref	klaus critique également le budget .
Src	and one has come from music .
Bsl	et un a de la musique .
Rdm	et on vient de musique .
PS	et on a viennent de musique .
All	et de la musique .
Infr	et un est venu de la musique .
Ref	et un vient du monde de la musique .

Figure 4: Examples of two translations for each of the SMT systems built: Src (source sentence), Bsl (baseline), Rdm (random selection), PS (probabilistic sampling), All (all the data available), Infr (Infrequent n-grams) and Ref (reference).

6 Discussion

Bilingual sentence selection (BSS) might be understood to be closely related to adaptation, even though both paradigms tackle problems which are, in essence, different. The goal of an adaptation technique is to *adapt model parameters*, which have been estimated on a large out-of-domain (or generic) data set, so that they are

best suitable for dealing with a domain-specific test set. This adaptation process is ought to be achieved by means of a (potentially small) adaptation set, which belongs to the same domain as the test data. In contrast, BSS tackles with the problem of how to *select samples* from a large pool of training data, regardless of whether such pool of data is in-domain or out-of-domain. Hence, in one case we can assume to have a fairly well estimated translation model, which is to be adapted, whereas in BSS we still have full control over the estimation of such model and need not to aim at a specific domain, although it might often be so.

BSS is related with *instance weighting* (Jiang and Zhai, 2007; Foster et al., 2010). Adaptation and BSS can be considered to be orthogonal (yet complementary) problems under the instance weighting paradigm. In such case, instance weighting can be considered to span a complete paradigmatic space between both. At one end, there is sample selection (BSS for SMT), while at the other end there is adaptation. For instance, it is quite common to confront the adaptation problem by extracting different phrase-tables from different corpora, and then interpolate such tables. This technique could be also applied to promote the performance of the system built by means of BSS. However, this is left out as future work.

We thoroughly analysed two BSS approaches that obtain competitive results, while using a small fraction of the training data, although there is still much to be gained. For instance, oracle results have also been reported in this work, yielding improvements of up to 10 BLEU points. Even though the use of an oracle typically implies that the results obtained are not realistic, recall that the proposed oracle is special, in the sense that it only uses the reference sentences for the specific purpose of selecting training samples, but the references are not included into the training data as such. This is useful for assessing the potential behind BSS: ideally, if we were able to design a BSS strategy that, without using the references, would select exactly those training samples, we would be boosting system performance by 10 BLEU points. This re-states BSS as a compelling technique that has not yet received the attention it deserves.

BSS is not aimed at optimising computational requirements, but does so as a byproduct. This may seem despicable but it would allow to run more experiments with the same resources, use larger corpora or even more complex techniques, such as synchronous grammars or hierarchical models. For instance, the infrequent n-grams technique has beaten all the other systems using just a small fraction of the corpus, only 0.5%, and is yet able to outperform a system trained with all the data by 0.9 BLEU points and the random baseline by 3 points. This baseline has been proved to be difficult to beat by other works.

Preliminary experiments were performed in order to analyse the perplexity of the references, the number of out of vocabulary words (OoVs) and the ratio of target-source phrases. These experiments revealed that the improvements obtained are largely correlated with a decrease in perplexity and in the number of OoVs. On the one hand, reducing the amount of OoVs was mirrored by an important improvement in BLEU when the amount of additional data was small, and also entailed a decrease in perplexity. However, a reduction in perplexity by itself did not always imply significant improvements. Moreover, no real conclusion could be drawn from the analysis of target-source phrase ratio. Hence, we understand that the improvements obtained are provided mainly by a more specialised estimation of the model parameters. However, further experiments should still be conducted in order to verify this conclusion.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement nr. 287755. This work was also supported by the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018), and iTrans2 (TIN2009-14511) project. Also supported by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project and Instituto Tecnológico de León, DGEST-PROMEP y CONACYT, México.

References

Sankaranarayanan Ananthakrishnan, Rohit Prasad, David Stallard, and Prem Natarajan. 2010. Discriminative sample selection for statistical machine translation. In *Proc. of the EMNLP*, pages 626–635, Cambridge, MA, October.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proc of the EMNLP*, pages 355–362.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1994. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc of the WSMT*, pages 1–28, Athens, Greece, March.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proc. of the MATR(ACL)*, pages 17–53, Uppsala, Sweden, July.

Sanjoy Dasgupta. 2009. The two faces of active learning. In *Proc. of The twentieth Conference on Algorithmic Learning Theory*, page 1, Porto (Portugal), October.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proc. of the EMNLP*, pages 451–459, Cambridge, MA, October.

Guillem Gascó, Vicent Alabau, Jesús Andrés-Ferrer, Jesús González-Rubio, Martha-Alicia Rocha, Germán Sanchis-Trilles, Francisco Casacuberta, Jorge González, and Joan-Andreu Sánchez. 2010. ITI-UPV system description for IWSLT 2010. In *Proc. of the IWSLT 2010*, Paris, France, December.

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *Proc. of HLT/NAACL'09*, pages 415–423, Morristown, NJ, USA.

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proc. of ACL'07*, pages 264–271.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. *Proc. of ICASSP*, II:181–184, May.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christie Moran, Richard Zens, Chris Dyer, Ontraj Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL*, pages 177–180.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Ann Irvine, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Ziyuan Wang, Jonathan Weese, and Omar Zaidan. 2010. Joshua 2.0: A toolkit for parsing-based machine translation with syntax, semirings, discriminative training and other goodies. In *Proc. of the MATR(ACL)*, pages 139–143, Uppsala, Sweden, July.

Yajuan Lu, Jin Huang, and Qun Liu. 2007. Improving statistical machine translation performance by training data selection and optimization. In *Proc. of the EMNLP-CoNLL*, pages 343–350, Prague, Czech Republic, June.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proc. of the EMNLP*, pages 708–717, Singapore, August.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *ACL (Short Papers)*, pages 220–224.

Franz J. Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of ACL*, pages 295–302.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29, pages 19–51.

Kishore Papineni, Salim Roukos, and Todd Ward. 1998. Maximum likelihood and discriminative training of direct translation models. In *Proc. of ICASSP'98*, pages 189–192.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: A method for automatic evaluation of machine translation. In *Technical Report RC22176 (W0109-022)*.

Michael Paul, Marcello Federico, and Sebastian Stker. 2010. Overview of the IWSLT 2010 evaluation campaign. In *Proc. of the IWSLT 2010*, Paris, France, December.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA'06*.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. of ICSLP*.

Elia Yuste, Manuel Herranz, Antonio Lagarda, Lionel Tarazón, Isaías Sánchez-Cortina, and Francisco Casacuberta. 2010. Pangeamt - putting open standards to work... well. In *Proc. of the AMTA2010*. Denver, CO, USA, November.