Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 101-108 recibido 26-03-2015 revisado 28-04-2015 aceptado 15-05-2015

An Empirical Analysis of Data Selection Techniques in Statistical Machine Translation *

Análisis empírico de técnicas de selección de datos en traducción automática estadística

Mara Chinea-Rios
Universitat Politècnica de València
Camino de Vera s/n, Valencia, Spain
machirio@prhlt.upv.es

Germán Sanchis-Triches
Universitat Politècnica de València
Camino de Vera s/n, Valencia, Spain
gsanchis@dsic.upv.es

Francisco Casacuberta
Universitat Politècnica de València
Camino de Vera s/n, Valencia, Spain
fcn@prhlt.upv.es

Resumen: La adaptación de dominios genera mucho interés dentro de la traducción automática estadística. Una de las técnicas de adaptación esta basada en la selección de datos que tiene como objetivo seleccionar el mejor subconjunto de oraciones bilingües de un gran conjunto de oraciones. En este artículo estudiamos como afectan los corpus bilingües empleados por los métodos de selección de frases en la calidad de las traducciones.
Palabras clave: traducción automática estadistica, adaptación dominios, selección de frases bilingües, n-gramas infrecuentes, entropía cruzada

Abstract: Domain adaptation has recently gained interest in statistical machine translation. One of the adaptation techniques is based in the selection data. Data selection aims to select the best subset of the bilingual sentences from an available pool of sentences, with which to train a SMT system. In this paper, we study how affect the bilingual corpora used for the data selection methods in the translation quality.
Keywords: statistical machine translation, domain adaptation, bilingual sentence selection, infrequent n-gram, cross-entropy

1 Introduction

Statistical machine translation (SMT) system quality depends on the available parallel training data. Two factors are important: the size of the parallel training data and the domain. A small set of training data leads to poorly estimated translation models and consequently poor translation quality. Unfortunately, we do not have parallel data in all domains. For this reason, the translation quality gets worse when we do not have enough training data for the specific domain we need to tackle in our test set. Intuitively, domain adaptation methods try to make a better use of the part of the training data that is more similar, and therefore more relevant, to the text that is being translated (Sennrich, 2013).

There are many domain adaptation methods that can be split into two broad categories. Domain adaptation can be done at the corpus level, for example, by weighting, selecting or joining the training corpora. In contrast, domain adaptation can also be done at the model level by adapting directly the translation or language models.

Bilingual sentence selection (BSS) aims to select the best subset of bilingual sentences from an available pool of sentence pairs. Here, we focus on studying the performance of two different BSS strategies. We will refer to the pool of sentences as *out-of-domain* (OoD) corpus because we assume that it belongs to a different domain than the one to be translated. We will refer to the corpus of the domain to be translated as *in-domain* (ID) corpus.

Since we will be analysing the BSS techniques as applied to the specific case of SMT, we review briefly the SMT (Papineni, Roukos, and Ward, 1998; Och and Ney, 2002) frame

* The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 287576 (CasMaCat). Also funded by the Generalitat Valenciana under grant Prometeo/2009/014.

ISSN 1135-5948

work: given an input sentence $\mathbf{x}$ from a certain source language, the purpose is to find an output sentence $\mathbf{y}$ in a certain target language such that:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{x}, \mathbf{y}) \qquad (1)$$

where λ_m is the weight assigned to $h_m(\mathbf{x}, \mathbf{y})$ and $h_m(\mathbf{x}, \mathbf{y})$ is a score function representing an important feature for the translation of $\mathbf{x}$ into $\mathbf{y}$, as for example the language model of the target language, a reordering model, or several translation models. The weights λ_m are normally optimised with the use of a development set. The most popular approach for adjusting λ_m is the one proposed in (Och, 2003), commonly referred to as *minimum error rate training* (MERT).

The main contribution of this paper is:

- An empirical analysis of two BSS techniques with different corpora.

This paper is structured as follows. Section 2 summarises the related work. Section 3 presents the two BSS techniques selected, namely, infrequent n-grams recovery and cross entropy selection. In Section 4, experimental results are reported. Conclusions and future work are presented in Section 5.

2 Related work

State-of-the-art BSS approaches rely on the idea of choosing those sentence pairs in the OoD training corpus that are in some way similar to an ID training corpus in terms of some different metrics.

The simplest instance of this problem can be found in language modelling, where perplexity-based selection methods have been used (Gao et al., 2002). Here, OoD sentences are ranked by their perplex-ity score. Another perplexity-based approach is presented in (Moore and Lewis, 2010), where cross-entropy difference is used as a ranking function rather than just perplexity, in order to account for normalization. We apply this criterion for the task of selecting training data for SMT systems

Different works use perplexity-related BSS strategies (Axelrod, He, and Gao, 2011; Rousseau, 2013). In Axelrod, He and Gao (2011), the authors used three methods based in cross-entropy for extracting a pseudo ID corpus. This pseudo ID corpus is used to train small domain-adapted SMT systems. In (Rousseau, 2013) the authors describe the *XenC* open source toolkit for data selection. *XenC* uses the two strategies described in (Gao et al., 2002) and (Moore and Lewis, 2010).

Two different approaches are presented in (Gascó et al., 2012): one based on approximating the probability of an ID corpus and another one based on infrequent n-gram occurrence. The technique approximating the probability relies on preserving the probability distribution of the task domain by wisely selecting the bilingual pairs to be used. Hence, it is mandatory to exclude sentences from the pool that distort the actual probability. The technique based in infrequent n-gram occurrence will be explained in detail in the next section.

Other works have applied information retrieval methods for BSS (Lü, Huang, and Liu, 2007), in order to produce different submodels which are then weighted. In that work, authors define the baseline as the result obtained by training only with the corpus that shares the same domain with the test. Afterwards, they claim that they are able to improve the baseline translation quality by adding new sentences retrieved with their method. However, they do not compare their technique with a model trained with all the corpora available.

3 Data selection methods

In this section we present the two techniques that we have selected for our work. The first strategy we used in this work is infrequent n-grams recovery. This strategy was presented in Gascó et al., (2012).

The second strategy, proposed in Moore and Lewis, (2010), is based in cross-entropy. This strategy is used in many different works (Axelrod, He, and Gao, 2011; Rousseau, 201); Schwenk, Rousseau, and Attik, 2012; Sennrich, 2012). In these works, the authors report good results when using this strategy, and has become a defacto standard in the SMT research community.

3.1 Infrequent n-grams recovery

The main idea underlying the infrequent n-grams recovery strategy consists in increasing the information of the ID corpus by adding evidence for those n-grams that have been

seldom observed in the ID corpus. The n-grams that have never been seen or have been seen just a few times are called *infrequent n-grams*. An n-gram is considered infrequent when it appears less times than a given infrequency threshold t. Therefore, the strategy consists on selecting from the OoD corpus the sentences which contain the most infrequent n-grams in the source sentences to be translated.

Let X be the set of n-grams that appears in the sentences to be translated and $\mathbf{w}$ one of them; let $N_{\mathbf{x}}(\mathbf{w})$ be the counts of $\mathbf{w}$ in a given source sentence $\mathbf{x}$ of the OoD corpus, and $C(\mathbf{w})$ the counts of $\mathbf{w}$ in the source language ID corpus. Then, the infrequency score $i(\mathbf{x})$ is defined as:

$$i(\mathbf{x}) = \sum_{\mathbf{w} \in X} \min(1, N_{\mathbf{x}}(\mathbf{w})) \max(0, t - C(\mathbf{w})) \quad (2)$$

Then, the sentences in the OoD corpus are scored using Equation 2. The sentence $\mathbf{x}^*$ with the highest score $i(\mathbf{x}^*)$ is added to the ID corpus and removed from the OoD sentences. The counts of the n-grams $C(\mathbf{w})$ are updated with the counts $N_{\mathbf{x}^*}(\mathbf{w})$ within $\mathbf{x}^*$ and therefore the scores of the OoD corpus are updated. Note that t will determine the maximum amount of sentences that can be selected, since when all the n-grams within X reach the t frequency no more sentences will be extracted from the OoD corpus.

3.2 Cross-entropy selection

As mentioned in Section 2, one established method consists in scoring the sentences in the OoD corpus by their perplexity score. We follow the procedure described in Moore and Lewis (2010), which uses the cross-entropy rather than perplexity. Perplexity and cross-entropy are monotonically related. The perplexity of a given sentence $\mathbf{x}$ with empirical n-gram distribution p given a language model q is:

$$2^{-\sum_x p(x) \log q(x)} = 2^{H(p,q)} \quad (3)$$

where $H(p, q)$ is the cross-entropy between p and q. The formulation proposed by Moore and Lewis (2010) is: Let I be an ID corpus and G be an OoD corpus. Let $H_I(\mathbf{x})$ be the cross-entropy, according to a language model trained on I, of a sentence $\mathbf{x}$ drawn from G. Let $H_G(\mathbf{x})$ be the cross-entropy of $\mathbf{x}$ according to a language model trained on G. The cross-entropy score of $\mathbf{x}$ is then defined as

$$c(\mathbf{x}) = H_I(\mathbf{x}) - H_G(\mathbf{x}) \quad (4)$$

In this work we will also analyse the effect of varying the order of the n-grams considered, since this will also imply that the final sentence selection will be different. Specifically, we will consider $n = \{2, 3, 4, 5\}$ grams.

4 Experiments

4.1 Experimental set-up

We evaluated empirically the methods described in the previous section. As ID data, we used two different corpora. The EMEA[1] corpus (Tiedemann, 2009) is available in 22 languages and contains documents from the European Medicines Agency. The other ID corpus is the News Commentary[2] (NC) corpus. The NC corpus is composed of translations of news articles. The main statistics of the ID corpora used are shown in Table 1. For the OoD corpora, we used two corpora belonging to different domains readily available in the literature. Table 2 shows the main features of the two OoD corpora. The Europarl[3] corpus is composed of translations of the proceedings of the European parliament (Koehn, 2005). The PatTR corpus[4] (Wäschle and Riezler, 2012) is a parallel corpus extracted from the MAREC patent collection.

All experiments were carried out using the open-source SMT toolkit Moses version phrase-based (Koehn et al., 2007). The language model used was a 5-gram, standard in SMT research, with modified Kneser-Ney smoothing (Kneser and Ney, 1995), built with the SRILM toolkit (Stolcke, 2002). The phrase table was generated by means of symmetrised word alignments obtained with GIZA++ (Och and Ney, 2003). The de-coder features a statistical log-linear model including a phrase-based translation model, a language model, a distortion model and word and phrase penalties. The log-lineal combi-nation weights in Equation 1 were optimized using MERT (minimum error rate training)(Och, 2003).

For each domain, we trained two differ-ent baselines with which to compare the systems obtained by data selection. The

[1] www.statmt.org/wmt14/medical-task/
[2] www.statmt.org/wmt13
[3] www.statmt.org/europarl/
[4] www.cl.uni-heidelberg.de/statnlpgroup/pattr/

Corpus		$\|S\|$	$\|W\|$	$\|V\|$
EMEA-Domain	EN	1.0M	12.1M	98.1k
	FR		14.1M	112k
Medical-Test	EN	1000	21.4k	1.8k
	FR		26.9k	1.9k
Medical-Mert	EN	501	9850	979
	FR		11.6k	1.0k
NC-Domain	EN	157k	3.5M	65k
	FR		4.0M	76k
NC-Test	EN	3000	56.0k	4.8k
	FR		61.5k	5.0k
NC-Mert	EN	2050	43.4k	3.9k
	FR		47.1k	4.1k

Table 1: ID corpora main figures. (EMEA-Domain and NC-Domain) are the ID corpora, (Medical-Test and NC-Test) are the evaluation data and (Medical-Mert and NC-Mert) are development set. M denotes millions of elements and k thousands of elements, $|S|$ stands for number of sentences, $|W|$ for number of words (tokens) and $|V|$ for vocabulary size (types).

Corpus		$\|S\|$	$\|W\|$	$\|V\|$
Europarl	EN	2.0M	50.2M	157k
	FR		52.5M	215k
PatTR	EN	3.4M	78.6M	190k
	FR		81.8M	212k

Table 2: OoD corpora main figures (See Table 1 for an explanation of the abbreviations).

first baseline was obtained by training the SMT system only with ID training data: EMEA-Domain and NC-Domain, obtaining the `baseline-emea` and `baseline-nc` baselines, respectively. The second baseline was obtained by training the SMT system with a concatenation of either of the OoD corpora (Europarl or PatTR) and the ID training data (EMEA-Domain or NC-Domain):

- `bsln-emea-euro`: EMEA ∪ Europarl
- `bsln-nc-euro`: NC ∪ Europarl
- `bsln-emea-pattr`: EMEA ∪ PatTR
- `bsln-nc-pattr`: NC ∪ PatTR.

Results are shown in terms of BLEU (Papineni et al., 2002), measures the precision of uni-grams, bigrams, trigrams, and four-grams with respect to a set of reference translations, with a penalty for too short sentences.

4.2 Results for the infrequent n-grams technique

In this section, we present the experimental results obtained by infrequent n-grams recovery for each set-up presented in Section 4.1.

Figures 1 and 2 show the effect of adding sentences using infrequent n-grams selection, up to the point where the specific value of t does not allow to select further sentences. In addition, the result obtained with the two baseline systems is also displayed. We only show results for threshold values $t = \{10, 20\}$ for clarity, although experiments were also carried out for $t = \{10, 15, 20, 25, 30\}$ and such results presented similar curves.

Figure 1 shows the principal result obtained using the Europarl OoD corpus. Several conclusion can be drawn:

- The translation quality provided by the infrequent n-grams technique is large better in term of BLEU than the results achieved with the system `baseline-nc` and `baseline-emea`.
- Selecting sentences with the infrequent n-grams technique provides better results than including the OoD corpus in the SMT system with Medical domain (`bsln-emea-euro`). Specifically, the improvements obtained are in the range 0.70 BLEU points using less than 4% of the Europarl OoD corpus. Different result are obtained with News domain. In this scenario, the infrequent n-grams technique does not provide significantly better results than including the OoD corpus in the SMT system with News domain (`bsln-nc-euro`). But the results are very similar using less than 8% of the OoD corpus.
- As expected, $t = 20$ allows to select more sentences than $t = 10$, which also leads to higher BLEU scores. The results with $t = 10$ are slightly worse than with $t = 20$, for the same amount of sentences. We understand that this is because $t = 20$ entails a better estimation of the n-grams considered infrequent.

Figure 2 shows the principal results obtained using PatTR OoD corpus.

- In this scenario, the results achieved by the baseline systems do not show a significant difference when including the

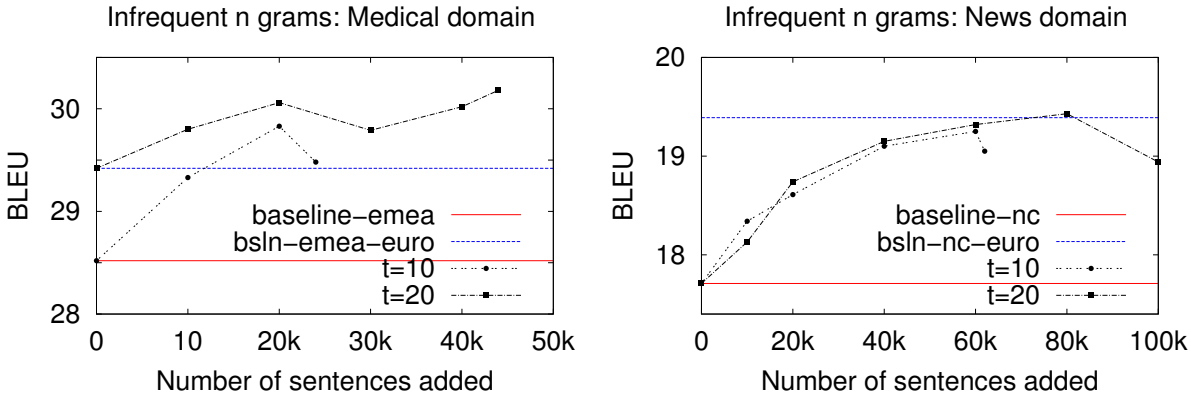

Figure 1: Effect over the BLEU score using infrequent n-grams recovery for two ID corpora EMEA and News and the Europarl OoD corpus. Horizontal lines represent the score when using the ID corpus and all the data available.

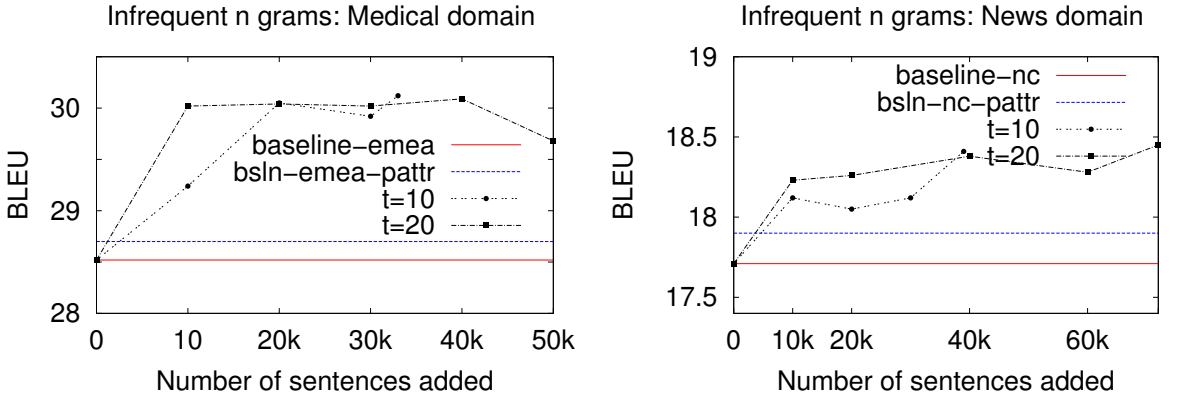

Figure 2: Effect over the BLEU score using infrequent n-grams recovery for two ID corpora EMEA and News and the PatTR OoD corpus. Horizontal lines represent the score when using the ID corpus and all the data available.

OoD PatTR data. We conclude that this corpus does not provide relevant information for the SMT system.

- The translation quality provided by the infrequent n-grams technique is large better in term of BLEU than the results achieved with all baseline systems, which evidence that the selection strategy is able to make a good use of the OoD data, even if such data as a whole does not seem to be useful. We understand that this is important, since it proves the utility of the BSS strategy.

4.3 Results for cross-entropy strategy

In this section, we present the experimental results obtained by the cross-entropy strategy for each set-up presented in Section 4.1.

Figures 3 and 4 show the effect of adding sentences by means of the cross-entropy strategy. We only show results using both 2-grams and 5-grams for clarity, although experiments were also carried out for $n = \{2, 3, 4, 5\}$.

- Adding sentences selected by means of cross-entropy improves over `baseline-emea` and `baseline-nc` from the very beginning, except the results obtained when testing in the medical domain and training with the PatTR OoD corpus.
- Cross-entropy data selection is not able to achieve improvements over training with all the data available when the Europarl corpus is considered as OoD corpus. When considering the PatTR, slight improvements are achieved, although such improvements are not very significant.
- In most cases, the order of the n-grams

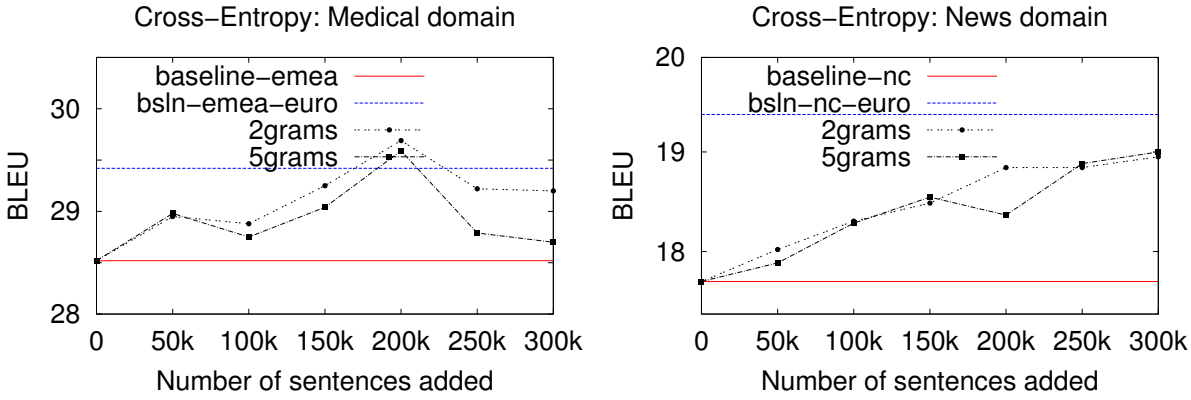

Figure 3: Effect to adding sentences over the BLEU score using cross-entropy strategy (with different n-gram value) for two ID corpora EMEA and News and the Europarl OoD corpus. Horizontal lines represent the score when using the ID corpus and all the data available.

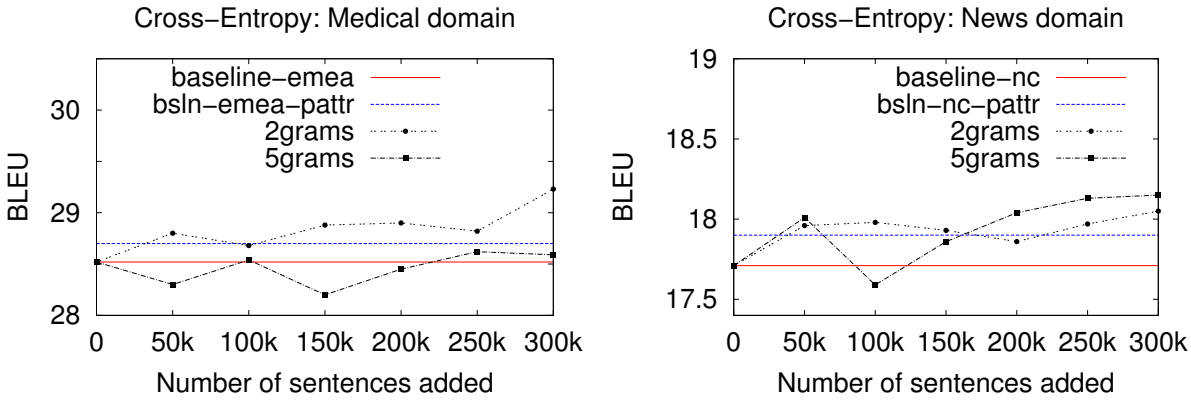

Figure 4: Effect to adding sentences over the BLEU score using cross-entropy strategy (with different n-gram value) for two ID corpora EMEA and News and the PatTR OoD corpus. Horizontal lines represent the score when using the ID corpus and all the data available.

used does not seem to affect significantly the translation quality, although using 2-grams provides slightly better results.

- Lastly, it is also worth noting that the results obtained with the cross-entropy strategy are slightly worse than the ones obtained with infrequent n-gram recovery in all the set-ups analysed, even though more sentences are considered when using cross-entropy.

4.4 Example Translations

Translation examples are shown in Table 3. In the first example, both the infrequent n-gram selection and baseline systems are able to obtain the sing % as appears in the reference. This in not only casual, since, by ensuring coverage for the infrequent n-grams only up to a certain t, we avoid distorting the specificities of the ID data. All the systems present the same lexical choice error with word (*développer*). However, this is so because this is the most likely translation in our data, both ID and OoD. The second example presents a sentence belonging to the NC-test set. None of the systems analysed achieved to produce the correct translation of "republican strategy". However, the "all" system did manage to produce the right reordering, even though the genre in *républicain* was not matched, and then word *á* was introduced instead of "pour". Note that, even if the cross-entropy translation of "counter" is different from the reference, it is semantically equivalent (even though BLEU would penalise it). This is, again, a lexical choice error.

4.5 Summary of the results

Table 4 shows the best results obtained with both strategies for each combination of the ID and OoD corpora. We can see the difference in number of selected sentences be-

Src	about 5 percent of people with ulcerative colitis develop colon cancer .
Bsl	environ 5 % des personnes avec colite ulcéreuse *de développer* un cancer du colon .
All	environ 5 *pour cent* des personnes avec colite ulcéreuse *développer* un cancer du colon .
Infr	environ 5 % des personnes avec colite ulcéreuse de *développer* un cancer du colon .
Entr	environ 5 *pour cent* des personnes avec colite ulcéreuse de *développer* un cancer du colon .
Ref	environ 5 % des personnes souffrant de colite ulcéreuse sont atteintes de cancer du côlon.
Src	a republican strategy to counter the re-election of obama
Bsl	une républicain stratégie pour contrer la réélection d'obama
All	une stratégie républicain á contrer la réélection d'obama
Infr	une républicain stratégie pour contrer la réélection d'obama
Entr	une républicain stratégie pour contrecarrer la réélection d'obama
Ref	une stratégie républicaine pour contrer la réélection d'obama

Table 3: Example of two translations for each of the SMT systems built: Src (source sentence), Bsl (baseline), All (all the data available), Infr (Infrequent n-grams), Entr (Cross-entropy) and Ref (reference).

Data	Strategy	BLEU	$\|S\|$
EMEA-Euro	ID	28.5	1.0M
	all data	29.4	1.0M+1.4M
	cross-entropy	29.7	1.0M+200k
	infreq. $t = 20$	30.2	1.0M+44k
EMEA-PatTR	ID	28.5	1.0M
	all data	28.7	1.0M+3.3M
	cross-entropy	29.2	1.0M+300k
	infreq. $t = 20$	30.2	1.0M+62k
NC-Euro	ID	17.7	117k
	all data	19.4	117k+1.4M
	cross-entropy	19.0	117k+300k
	infreq. $t = 20$	19.4	117k+80k
NC-PatTR	ID	17.7	117k
	all data	17.9	117k+3.3M
	cross-entropy	18.2	117k+300k
	infreq. $t = 20$	18.5	117k+72k

Table 4: Summary of the best results obtained with each setup. `Euro` stands for Europarl and $|S|$ for number of sentences, which are given in terms of the ID corpus size, and (+) the number of sentence selected.

tween infrequent n-grams and cross entropy. The cross-entropy strategy selects more sentences and the results achieved are worse than when using infrequent n-grams. We understand that this is because the infrequent n-grams technique selects more relevant sentences from the OoD corpus.

We observe performance differences between both ID corpora (EMEA and NC). Results obtained with the NC corpus seem to indicate that it is not an adequate corpus for testing adaptation techniques, as observed in our results and also in related work (Haddow and Koehn, 2012; Irvine et al., 2013). Hence, we disrecommend using the NC corpus for adaptation experiments, as it might lead to misleading results.

5 Conclusions and future work

Data selection has been receiving an increasing amount of interest within the SMT research community. In this work, we study the effect of using different data sets with two popular BSS strategies. The results obtained are similar in term of BLEU, although the best results were obtained by infrequent n-grams. Such conclusion is coherent across all combinations of corpora studied, i.e., ID and OoD. Finally, the BSS techniques obtain positive results using only a small fraction of the training data. These results show the importance of the data sets used: it is important to use a general-domain OoD corpus, such Europarl. Moreover, the NC corpus is not appropriate corpus to be used for the evaluation of domain adaptation methods.

In future work, we intend to combine the two strategies proposed and will develop new experiments with bigger and more diverse data sets. In addition, we will also study different data selection methods using a vectorial representation of sentences.

References

Axelrod, A., X. He, and J. Gao. (2011). Domain adaptation via pseudo in-domain data selection. In *Proc. of the EMNLP*, pages 355–362.

Gao, J., J. Goodman, M. Li, and K. Lee.

(2002). Toward a unified approach to statistical language modeling for chinese. *ACM TALIP*, 1:3–33.

Gascó, G., M.A. Rocha, G. Sanchis-Trilles, J. Andrés-Ferrer, and F. Casacuberta. (2012). Does more data always yield better translations? In *Proc. of the EACL*, pages 152–161.

Haddow, B. and P. Koehn. (2012). Analysing the effect of out-of-domain data on smt systems. In *Proc. of the Seventh Workshop on Statistical Machine Translation*, pages 422–432.

Irvine, A., J. Morgan, M. Carpuat, H. Daumé III, and D. Munteanu.(2013). Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics*, 1:429–440.

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

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, pages 79–86.

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

Lü, Y., J. Huang, and Q. Liu. (2007). Improving statistical machine translation performance by training data selection and optimization. In *Proc. of the EMNLP-CoNLL*, pages 343–350.

Moore, R. C. and W. Lewis. (2010). Intelligent selection of language model training data. In *Proc. of the ACL*, pages 220–224.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proc. of the ACL*, pages 160–167.

Och, F. J. and H. Ney. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proc. of the ACL*, pages 295–302.

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

Papineni, K., S. Roukos, T. Ward, and W. J. Zhu. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proc. of the ACL*, pages 311–318.

Papineni, K. A, S. Roukos, and R. T. Ward. (1998). Maximum likelihood and discriminative training of direct translation models. In *Proc. of the International Conference on Acoustics Speech and Signal Processing*, pages 189–192.

Rousseau, A. (2013). Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, 100:73–82.

Schwenk, H., A. Rousseau, and M. Attik. (2012). Large, pruned or continuous space language models on a gpu for statistical machine translation. In *Proc. of the NAACL*, pages 11–19.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proc. of the EACL*, pages 539–549.

Sennrich, R. (2013). *Domain adaptation for translation models in statistical machine translation.* Ph.D. thesis, University of Zurich.

Stolcke, A. (2002). Srilm-an extensible language modeling toolkit. In *Proc. of the Seventh International Conference on Spoken Language Processing.*

Tiedemann, J.(2009). News from opus- a collection of multilingual parallel corpora with tools and interfaces. In *Proc. of the Recent advances in natural language*, pages 237–248.

Wäschle, K. and S. Riezler. (2012). Analyzing Parallelism and Do-main Similarities in the MAREC Patent Corpus. *Multidisciplinary Information Retrieval*, pages 12–27.