Data selection for NMT using Infrequent n-gram Recovery

Zuzanna Parcheta[1] **Germán Sanchis-Trilles**[1] **Francisco Casacuberta**[2]
[1]Sciling S.L., Carrer del Riu 321,Pinedo, 46012, Spain
[2]PRHLT Research Center, Camino de Vera s/n, 46022 Valencia, Spain
{zparcheta, gsanchis}@sciling.com
fcn@prhlt.upv.es

Abstract

Neural Machine Translation (NMT) has achieved promising results comparable with Phrase-Based Statistical Machine Translation (PBSMT). However, to train a neural translation engine, much more powerful machines are required than those required to develop translation engines based on PBSMT. One solution to reduce the training cost of NMT systems is the reduction of the training corpus through data selection (DS) techniques. There are many DS techniques applied in PBSMT which bring good results.

In this work, we show that the data selection technique based on infrequent n-gram occurrence described in (Gascó et al., 2012) commonly used for PBSMT systems also works well for NMT systems. We focus our work on selecting data according to specific corpora using the previously mentioned technique. The specific-domain corpora used for our experiments are IT domain and medical domain. The DS technique significantly reduces the execution time required to train the model between 87% and 93%. Also, it improves translation quality by up to 2.8 BLEU points. The improvements are obtained with just a small fraction of the data that accounts for between 6% and 20% of the total data.

1 Introduction

Until recently, machine translation (MT) systems were based mostly on PBSMT. Today, the state of the art of MT is NMT. It has been shown that neural networks can improve the quality of translations by up to several BLEU points and also make them more fluid (Toral and Sánchez-Cartagena, 2017). However, NMT is computationally much more expensive. To train an NMT engine, much more powerful machines are required than would be used for building translation engines based on PBSMT. For example, NMT engines require more RAM memory, one or several GPUs and storing the models requires more storage capacity. Also, the training time of an NMT system is significantly longer than that of the systems based on PBSMT (Shterionov et al., 2017). One solution to reduce the training cost of NMT systems is the reduction of the training corpus through DS techniques. Bilingual sentence selection (BSS) is a type of DS where the best subset of bilingual sentences from the available parallel corpora is selected and leveraged to train a translation system. To date, many DS techniques are known that are applied to PBSMT systems, bringing very promising results. Some of them not only reduce the training time but also outperform a system where all the bilingual data available is used, given that the selected sentences are better suited to the domain being dealt with.

In this work, we demonstrate that a DS technique commonly used for PBSMT can also yield satisfactory results when applied in NMT systems. To prove a good performance of DS in NMT we select sentences from a large amount of data from different domains with the purpose of

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 219–227
Alacant, Spain, May 2018.

enlarging a small size, in-domain training corpus. The selection of more suitable sentences achieves improvements in translation quality.

2 Related Work

When creating a machine translation system, it is important to select high-quality bilingual data with a domain similar to the one in which the translation system will be used.

There are multiple techniques of DS for PBSMT based on perplexity as (Gao et al., 2002), where the authors use maximum-likelihood based methods to select the lexicon, segment words, filter and adapt the training data, and reduce language model size. In (Moore and Lewis, 2010), data selection is done comparing the cross-entropy according to domain-specific and non-domain specific. In (Axelrod et al., 2011), sentences are selected with a bilingual cross-entropy based method. The selected subset is used to train a small domain-adapted PBSMT system. This domain-adapted system is combined with the real in-domain PBSMT system.

Also, there are techniques based on distributed representations of words. In (Chen et al., 2016) and (Chen and Huang, 2016), sentences are selected using a convolutional neural network. In (Chinea-Rios et al., 2016), a continuous vector-space representation of word sequences is used for selecting the best subset of a bilingual corpus. In (Peris et al., 2017), a new data selection method is developed, based on a neural network classifier.

Other data selection techniques rely on information retrieval based methods. In (Lu et al.), training data is adapted by redistributing the weight of each training sentence pair.

There are also DS techniques which select sentences relying on information from the development and test set. In (Gascó et al., 2012) two data selection techniques are presented: 1) Probabilistic sampling, that introduces new sentences into the in-domain corpus without distorting the original distribution. First, the sentences are selected according to length, then according to probability. The second technique presented in that work is infrequent n-gram recovery. This technique relies on the idea of enforcing model coverage for those n-grams that are present in the (source) test set. In (Biçici and Yuret, 2011), the authors explore the use of a data selection in a transductive scenario. Feature decay algorithms increase the diversity of the training set by devaluing features that are already included.

All commented techniques were initially implemented for PBSMT systems. There are also some techniques designed explicitly for NMT systems. In (Farajian et al., 2017), the authors present an instance-based adaptive NMT approach that effectively handles translation requests from multiple domains in an unsupervised manner, that is without knowing the domain labels. In (Chinea-Rios et al., 2017), the method developed consists in selecting, from a large monolingual pool of sentences in the source language, those instances that are more related to a given test set. Next, this selection is automatically translated and the general neural machine translation system is fine-tuned with this data.

Also, there are some works that compare the effectiveness of data selection techniques in PBSMT and NMT. In (van der Wees et al., 2017), the authors compare the effects of a commonly used data selection approach (bilingual cross entropy) on PBMT and NMT using four different domains. They also introduce dynamic data selection as a way to make data selection profitable for NMT.

3 Infrequent n-gram Recovery

The data selection technique used in this work is called Infrequent n-gram Recovery (Gascó et al., 2012). The main use of this technique is when the in-domain corpus provided is too small to train properly the translation engine. This technique consists on enlarging the in-domain training set by selecting sentences from a non domain-specific pool of sentences to maximise the coverage of n-grams which appear in the test and development set. For this, it is necessary to establish the minimum number of occurrences (t) required for a certain n-gram to be considered as infrequent, and also the order n of the n-grams (unigrams, bigrams, 3-grams etc.) that will be considered. The selected sentences will contain n-grams considered infrequent. With that we ensure that the training set will contain all n-grams from test and development set t times, as long as this is possible with the available out of domain dataset. The pool of sentences will be oppositely denoted as the *out-of-domain* corpus.

Sentences in the out-of-domain pool are sorted by their infrequency score in order to select first the sentences which most improve the coverage of n-grams belonging to the in-domain dataset which might be considered infrequent. Let χ be the set of *n*-grams that appear in the sentences to be translated and **w** one of them; $C(\mathbf{w})$ the counts of **w** in the source language training set; t the threshold of counts when an *n*-gram is considered infrequent, and $N(\mathbf{w})$ the counts of **w** in the source sentence **f** to be scored. The infrequency score of **f** is:

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

It already was demonstrated that the Infrequent n-gram Recovery technique works very well in PBSMT systems improving up to 1 point of BLEU when compared to training with all the data available (in-domain + out-of-domain), while using only 0.5% of total data. The fact, that the Infrequent n-gram Recovery technique works well in PBSMT system does not mean that it will work fine for NMT, since PBSMT and NMT build the translation model in very different ways. PBSMT splits sentences into smaller chunks and looks for similar occurrences in other languages according to a statistical model. The alignment matrix can not be well estimated if words and n-grams appear rarely in the training corpus. Also, the out-of-vocabulary words can not be translated by PBSMT model. The behaviour of NMT systems is different to PBSMT. NMT generates sequence of words in the target language given an input sequence of words in the source language. The translation is done following an encoder–decoder architecture. The encoder represents the input sequence using a word embedding model (Mikolov et al., 2013), and the decoder generates the sentence in the target language word by word (Sutskever et al., 2014). In NMT, it is necessary to adjust hyper-parameters as learning rate, number of hidden layers, and number of epochs. NMT needs to deal with millions of parameters coming from each neural network unit (weights and biases) to adjust the translation model. The best model is then selected according to translation quality on the development set.

Up until now there is no study about the efficiency of Infrequent n-gram Recovery in NMT.

4 Experiments

The experiments were conducted using the OpenNMT (Klein et al., 2017) deep learning framework based in Torch. This toolkit is mainly specialised in sequence-to-sequence models covering a variety of tasks such as machine translation, image to text, and speech recognition.

All experiments were conducted using an NVIDIA GTX 1080 GPU with 8GB of RAM.

To select domain-specific sentences, we need a small size in-domain dataset and an out-of-domain dataset which contains sentences from different domains. Then, we select sentences from the out-of-domain corpus to enlarge the in-domain corpus.

4.1 Experimental setup

We used two in-domain corpora for our experiments: Medical Web Crawl and IT. Medical Web Crawl is a subset of the UFAL Medical Corpus[1], which contains specific medical vocabulary and expressions; the IT corpus[2] contains sentences belonging to the IT domain. Main figures of both corpora are shown in Tables 1 and 2.

Table 1: Medical Web Crawl main figures. k denotes thousands of elements and M denotes millions 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	130k	1.9M	44.0k
	Spanish	130k	2.1M	54.5k
dev	English	806	12.3k	2.9k
	Spanish	806	13.4k	3.5k
test	English	810	12.1k	2.8k
	Spanish	810	13.3k	3.3k

We use two different out-of-domain corpora for each in-domain corpus. In the case of the IT corpus we use Europarl[3] as the out-of-domain dataset. In the case of Medical Web Crawl, we use JRC-Acquis[4] and Europarl jointly. JRC-Acquis is a collection of legislative text of the European Union and Europarl is a parallel corpus extracted from the European Parliament website. The purpose of using two different corpora for each

[1] https://ufal.mff.cuni.cz/ufal_medical_corpus
[2] http://www.statmt.org/wmt16/it-translation-task.html
[3] http://opus.nlpl.eu/Europarl.php
[4] http://opus.nlpl.eu/JRC-Acquis.php

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

Subset	language	$\|S\|$	$\|W\|$	$\|V\|$
train	English	147.9k	1M	44.4k
	Spanish	147.9k	1M	50.3k
dev	English	1.7k	32.4k	2.9k
	Spanish	1.7k	34k	3.4k
test	English	857	15.6k	2k
	Spanish	857	17.4k	2.4k

domain was to analyse system performance under different conditions: 1) a first condition (IT domain) in which training the system on all the available data (in-domain and out-of-domain data) leads to better results than training it only on the in-domain data; and 2) a second experiment (medical domain) in which training the system on all the available data leads to worse results than training the system on only the in-domain data. These two different scenarios allow us investigate the behaviour of the DS selection technique used in this work in a scenario where similar-domain data is abundant, but also in a scenario where similar-domain data is scarce. In both cases, sentences longer than 40 words were pruned. Main figures of the out-of-domain corpora are shown in Table 3. All data was previously tokenised and lowercased.

Table 3: Out-of-domain corpora main figures. k denotes thousands of elements. $|S|$ stands for number of sentences and M denotes millions of elements., $|W|$ for number of running words, and $|V|$ for vocabulary size.

Corpus	language	$\|S\|$	$\|W\|$	$\|V\|$
Europarl	English	1.7M	32.8M	118k
	Spanish	1.7M	33.9M	167k
JRC +	English	2.2M	41.2M	151k
Europarl	Spanish	2.2M	43M	198k

We conducted data selection experiments using the Infrequent n-gram Recovery technique. For each in-domain dataset (Medical and IT), the experiments were performed considering n-grams with $n \in \{1, \ldots, 5\}$ for computing the infrequency score (Equation 1). For each n-gram we conducted experiments for thresholds $t \in \{10, 20, 30, 40\}$. The count of infrequent n-grams was done on test and development set jointly. The reason for doing so was that the best model in NMT is chosen according to the best BLEU achieved on the development set. To ensure similar conditions in development and in test, it is important to ensure that all n-grams from the test and development sets appear in the training set. The data selected, together with the in-domain corpus, were used to train the reduced model.

We trained a Byte Pair Encoding model (BPE) (Sennrich et al., 2015) on the selected data and we applied the BPE model to training, development and test set. Then, we trained a recurrent neural network (RNN) (Schuster and Paliwal, 1997) with long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) on encoder and decoder side, each of them with only one layer because of the high computational cost entailed. We used a global attention layer to improve translation by selectively focusing on parts of the source sentence during translation. We also used a dropout rate of 0.2, and the *adam* (Kingma and Ba, 2014) optimiser with learning rate of 0.0002. The model featured 512 hidden units and 512-dimensional embedding vectors. The training procedure was run for 40 epochs and we selected the best epoch according to the development set.

We considered three different baseline systems against which to compare our DS systems: first, a model trained only with in-domain data; second, a model trained with all data available (in-domain and out-domain corpora jointly); third, a model trained on data selected at random. For this last baseline, we repeated the random selection procedure 5 times, reporting in our experiments the average of those 5 different experiments

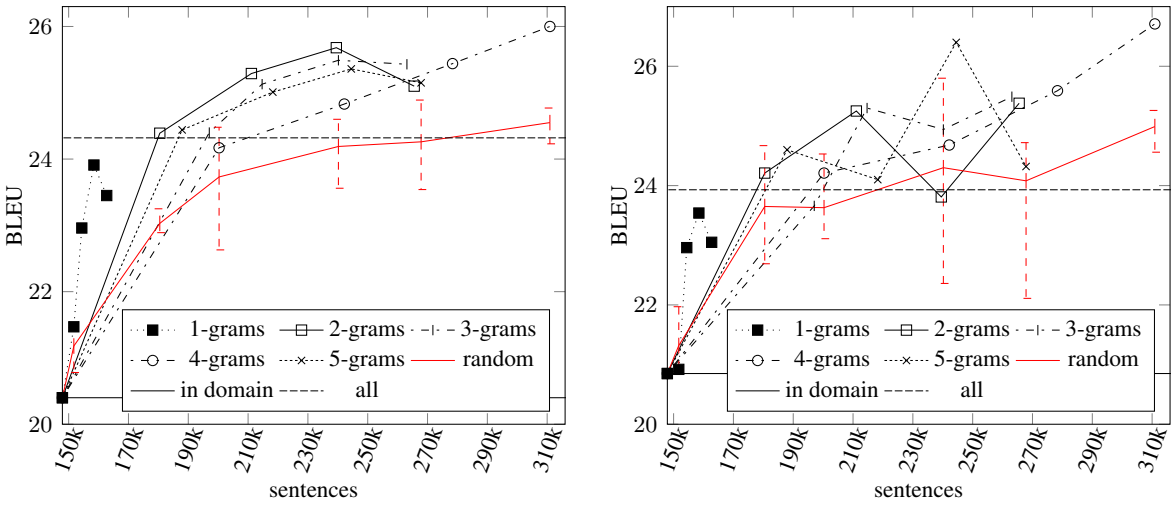
System performance was measured in terms of BLEU (Papineni et al., 2002), which measures n-gram precision with respect to a reference set, with a penalty for sentences that are too short.

4.2 Results

In this section we will analyse the results obtained for both domains. Given that the purpose of evaluating on the IT and medical domain is different, we will analyse the results obtained separately.

4.2.1 IT domain results

In Figure 1, we show BLEU scores for models trained with data selected according to different



(a) IT corpus development set

(b) IT corpus test set

Figure 1: Effect of adding sentences over the BLEU score in IT domain for n-grams $N = \{1, 2, 3, 4, 5\}$ with threshold $t = \{10, 20, 30, 40\}$, where $t = 10$ includes the lowest number of sentences. Figure 1a shows BLEU score for development set and Figure 1b shows BLEU score for test set. Red dashed lines show confidence intervals for random selection.

Table 4: Examples of translated sentences by the best model: 4-grams and t=40. In each example, we show source sentence (src), target sentence (ref), a hypothesis generated by the best model (hyp) and also, a hypothesis with a random model (hyp random). The random model is one of 5 random experiments conducted with the same number of sentences as our best model. This random model was chosen by BLEU score nearest to medium score from all 5 random models.

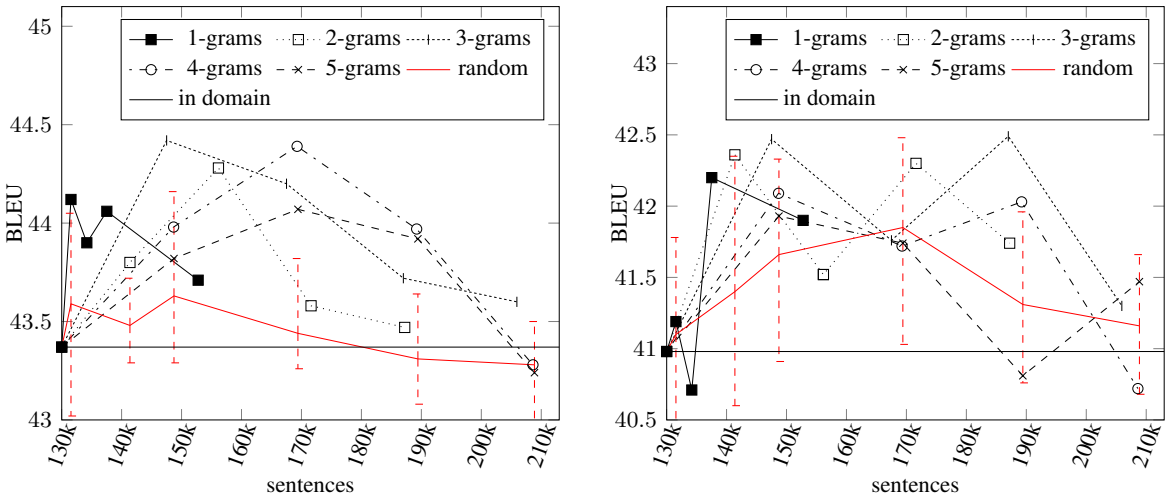
Example 1	
src	try to close and reopen the program.
ref	intente cerrar y abrir de nuevo el programa.
hyp	intentar cerrar y reabrir el programa.
hyp random	intentar cerrar y reabrir el programa.
Example 2	
src	try to shut down your computer, wait a few seconds, and boot it up again.
ref	intente apagar su ordenador, espere unos segundos, y reinícielo de nuevo
hyp	intentar cerrar su equipo, espere unos segundos, y la arranque de nuevo.
hyp random	intentar cerrar su equipo , espere unos pocos segundos y su arranque de nuevo.
Example 3	
src	click the apple icon, then select shut down.
trg	haga clic en el icono de apple y seleccione apagar.
hyp	haga clic en el icono de apple, luego seleccione cierre.
hyp random	pulse en el icono de apple cerrar.
Example 4	
src	someone probably reported you for copyright infringement.
trg	es probable que alguien haya informado al servicio de la infraccin de copyright.
hyp	alguien ha informado probablemente por infraccin de derechos de autor.
hyp random	alguien probablemente ha informado de sus derechos de autor.

order of n-grams and different threshold t. Also, we include the score obtained by a model trained only with in-domain data, and the score obtained by a model trained with all available data. Moreover, we show the average score of all 5 random models, with confidence intervals.

The best model obtained for the IT domain, according to the development set, is the model trained with data selected with n-grams up to order 4, with t=40. Our best model, obtained after epoch 7, reaches 26.7 BLEU on the test set. As described in Section 4.1, we compare our system against three different baselines:

1) Only in-domain data: The model trained only with in-domain data achieves 20.9 BLEU on the test set. Our system is able to improve this score by 5.8 BLEU points.

2) All data: The model trained with all data (in-domain and out-of-domain jointly) achieves 23.9 BLEU. Our system is able to improve this score by 2.8 BLEU points.



(a) Medical corpus development set

(b) Medical corpus test set

Figure 2: Effect of adding sentences over the BLEU score in medical domain for n-grams $N = \{1, 2, 3, 4, 5\}$ with threshold $t = \{10, 20, 30, 40\}$, where $t = 10$ includes the lowest number of sentences. Figure 2a shows BLEU score for development set and Figure 2b shows BLEU score for test set. Red dashed lines show confidence intervals for random selection.

Table 5: Examples of translated sentences by the best model: 3-grams and t=10. In each example, we show source sentence (src), target sentence (ref), a hypothesis generated by the best model (hyp) and also, a hypothesis with a random model (hyp random). The random model is one of 5 random experiments conducted with the same number of sentences as our best model. This random model was chosen by BLEU score nearest to medium score from all 5 random models.

Example 1	
src	wash hands and arms thoroughly after cleaning aquariums . or , wear rubber gloves when cleaning
ref	lávese muy bien las manos y los brazos después de limpiar acuarios o utilice guantes de caucho al realizar la limpieza .
hyp	lávese bien las manos y los brazos completamente después de limpiar los acuarios de limpieza o, use guantes de goma al limpieza.
hyp random	lávese bien las manos y los brazos bien después de limpiar los guantes de venta libre .
Example 2	
src	mellaril overdose ; hydrochloride - thioridazine overdose
ref	sobredosis de mellaril ; sobredosis de hidrocloruro de tioridazina
hyp	sobredosis de mogyil ; sobredosis de troridazina
hyp random	sobredosis de molcio
Example 3	
src	histamine h2 receptor blockers
trg	bloqueadores de los receptores h2 de la histamina .
hyp	bloqueadores h2 de la histamina los receptores de la histamina
hyp random	bloqueadores de los 2 bloqueadores
Example 4	
src	more than 200,000 had to go to the emergency room
trg	más de 200,000 acudieron a salas de emergencias,
hyp	más de 200.000 acudieron a la sala de urgencias
hyp random	más de 9,000 se sometieron a la sala de urgencias

3) Random selection: The average of scores achieved by the 5 random selections on the out-of-domain corpus is 25.0 points of BLEU. Our best model is able to improve this score by about 1.7 points of BLEU. Also, our model is also able to improve over the best of the models obtained with random selection by 1.4 points of BLEU.

The results described above are promising, since we are able to reach improvements in translation quality by selecting only 163k sentences, which represents 20% of all data available, and reduction of training size also implies reduction in model size and execution time: training a model with all the data available takes 10 days 6 hours, compared to 33 hours for the model trained with selected data using 4-grams and t=40, which implies a reduction of computational time by 87%.

Analysing the random selection score in

Figure 1a, it can be seen that the more sentences added in the random setting, the better the score in development. However, this is not so clear in test conditions (1b). In the case of the test set, the plot shows much more noise in the case of random selection.

Examples of translations generated by our model are shown in the Table 4. To compare the quality of the translations generated we show source and target sentences, which correspond to the reference translation. Also, we include the translation obtained by the best random selection, with a comparable number of selected sentences. We can see that the hypotheses of our model (hyp) and the hypothesis of the random model (hyp random) are pretty similar. In Example 1, both hypotheses are the same, and they are perfectly understandable synonyms of the reference translation. In Example 2, the hypotheses of random selection is mostly correct, but the use of the wrong article makes it difficult to understand. In Examples 3 and 4, our model generates a perfect translation. In contrast, the hypotheses generated by the random model have missed words in some cases, and in other present word substitutions that imply that the translation is disfluent and sometimes unable to convey the appropriate meaning.

4.2.2 Medical domain results

In Figure 2, we show the results for the medical domain. The score achieved by the model trained with all data is much lower than the score of the model trained only on in-domain data. For this reason, and for clarity purposes, we did not include the score of the system trained on all the data available. As in the case of the IT domain, we include the score of different systems obtained by selecting data with different order of n-grams and different thresholds t. Moreover, we show the average score of all 5 random selections, with confidence intervals.

In the case of the medical domain, the model trained only with in-domain data achieves 41 BLEU and leads to improvements over the model trained on all the data available, which reaches only 35 BLEU. It supports the hypothesis from (Gascó et al., 2012) that more data not always yields better results.

In case of the medical domain, the best model is trained on a selection obtained by n-grams up to order 3, with threshold t=10, after 19 epochs. This model achieves 42.5 BLEU on the test set with only 41.6K sentences added, which represents only 6% of all data. Our model achieves the following improvements over each of the three baselines described:

1) 1.5 points of BLEU over in-domain
2) 7.5 points of BLEU over out-of-domain
3) 0.8 points of BLEU over the average of scores of the 5 models trained with randomly selected data. Also, the system trained with Infrequent n-grams also improves by 0.1 BLEU over the best system obtained with random selection.

Observing the random-selection curve in Figure 2a, we realise that the more sentences added at random, the worse the BLEU score. We understand this is an evidence that signals that including sentences from an out-of-domain corpus leads to having the in-domain information overwhelmed, yielding a model which is not well suited for the specific domain at hand.

In the case of the development set, BLEU tends to degrade as soon as we add sentences after threshold *t=10* or *t=20*. However, in the case of the test set (Figure 2b), the plot is very noisy, and no clear pattern can be observed, both in the case of random selection and in the case of Infrequent n-gram selection.

It must be noted that training the system on all the data available took 12 days. In contrast, training the system with the selected data only took 17 hours, which entails a reduction of 93%. In Table 5, we show some examples of translations generated by our best model (3-grams with threshold *t=10*). Although a lot of sentences translated by our model and by the random model are very similar, we find some differences which lead us to think that our model generates better quality translations. In Examples 1, 2 and 3, shown in Table 5, the translations generated by random selection present some disfluencies. This model reorders and misses words causing the sentences to not be understandable. In Example 4, we can see that random selection translates a number incorrectly.

5 Conclusions

PBMT and NMT estimate the translation model in a different way. PBMT estimates the parameters

using statistical models and use word alignments to generate the translation. Instead, NMT features an encoder-decoder architecture. The encoder represents a sequence of input words mapping them to vectors of real numbers and then the decoder generates the output sequence in a word-by-word basis.

In our work, we show that Infrequent n-gram Recovery brings very satisfactory results when applied to NMT. We demonstrate that, by selecting a subset of data more suitable to a specific in-domain corpus, we can get a model whose quality can improve the quality of a model trained with all the data available (in-domain and out-of-domain data jointly). Such was the case with the IT corpus. In contrast, a less usual case is when the model trained with all data performs worse than one trained with only in-domain data. This was the case with the medical domain dataset. It can be due to very specific vocabulary appearing in the in-domain corpus, and such vocabulary not being frequent in the out-of-domain data. This entails that including sentences from different domains lead to worse translation quality. Despite this fact, the technique described manages to select only sentences that lead to improvements over the translation quality achieved by a system trained only with in-domain data.

In our experiments, we achieve improvements of up to 1.7 BLEU points over a model trained with a random selection of data. In the case of the IT corpus, we improved translation quality by about 2.8 points of BLEU when compared to a model trained on all the data available.

Another important issue is the reduction of execution time. By reducing the amount of training data, we achieved a reduction in execution time between 87% and 93%. We understand that this reduction is very important in the case of NMT, since training an NMT system can take up to several weeks. We demonstrate that with adequate DS, we can reduce execution time from 11 days to 17 hours, while simultaneously improving the translation quality achieved by a model trained with all the data available.

Acknowledgments

Work partially supported by MINECO under grant DI-15-08169 and by Sciling under its R+D programme.

References

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

Biçici, E. and Yuret, D. (2011). Instance selection for machine translation using feature decay algorithms. In *Proc. of WMT*, pages 272–283. ACL.

Chen, B. and Huang, F. (2016). Semi-supervised convolutional networks for translation adaptation with tiny amount of in-domain data. In *Proc. of SIGNLL-CoNLL*, pages 314–323.

Chen, B., Kuhn, R., Foster, G., Cherry, C., and Huang, F. (2016). Bilingual methods for adaptive training data selection for machine translation. In *Proc. of AMTA*, pages 93–106.

Chinea-Rios, M., Peris, A., and Casacuberta, F. (2017). Adapting neural machine translation with parallel synthetic data. In *proc. of WMT*, pages 138–147.

Chinea-Rios, M., Sanchis-Trilles, G., and Casacuberta, F. (2016). Bilingual data selection using a continuous vector-space representation. In *Proc. of SPR-SSPR*, pages 95–106. Springer.

Farajian, M. A., Turchi, M., Negri, M., and Federico, M. (2017). Multi-domain neural machine translation through unsupervised adaptation. In *proc. of WMT*, pages 127–137.

Gao, J., Goodman, J., Li, M., and Lee, K.-F. (2002). Toward a unified approach to statistical language modeling for chinese. *ACM*, pages 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 *Proc. of EACL*, pages 152–161.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, pages 1735–1780.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint*, arXiv:1412.6980.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). Opennmt: Open-source toolkit for neural machine translation. *arXiv preprints*, arXiv:1701.02810.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprints*, arXiv:1310.4546.

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

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

Peris, Á., Chinea-Ríos, M., and Casacuberta, F. (2017). Neural networks classifier for data selection in statistical machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):283–294.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprints*, arXiv:1508.07909.

Shterionov, D., Nagle, P., Casanellas, L., Superbo, R., and ODowd, T. (2017). Empirical evaluation of nmt and pbsmt quality for large-scale translation production. In *Proc. of EAMT*, pages 75–80.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proc. of NIPS*, volume 27, pages 3104–3112.

Toral, A. and Sánchez-Cartagena, V. M. (2017). A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. *arXiv preprint arXiv:1701.02901*.

van der Wees, M., Bisazza, A., and Monz, C. (2017). Dynamic data selection for neural machine translation. *arXiv preprint arXiv:1708.00712*.