Adaptation of Machine Translation Models with Back-Translated Data Using Transductive Data Selection Methods

Alberto Poncelas(✉), Gideon Maillette de Buy Wenniger, and Andy Way

ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
{alberto.poncelas,gemdbw}@gmail.com, andy.way@adaptcentre.ie

Abstract. Data selection has proven its merit for improving Neural Machine Translation (NMT), when applied to authentic data. But the benefit of using synthetic data in NMT training, produced by the popular back-translation technique, raises the question if data selection could also be useful for synthetic data?

In this work we use Infrequent n-gram Recovery (INR) and Feature Decay Algorithms (FDA), two transductive data selection methods to obtain subsets of sentences from synthetic data. These methods ensure that selected sentences share n-grams with the test set so the NMT model can be adapted to translate it.

Performing data selection on back-translated data creates new challenges as the source-side may contain noise originated by the model used in the back-translation. Hence, finding n-grams present in the test set become more difficult. Despite that, in our work we show that adapting a model with a selection of synthetic data is an useful approach.

Keywords: Data selection · Back-translation · Synthetic data · Neural machine translation · Feature decay algorithms · Infrequent n-gram recovery

1 Introduction

Neural Machine Translation (NMT) models tend to perform better with larger amounts of data. However, a smaller model trained with data in the same domain as the document to be translated (test set) may perform better than a bigger general-domain model.

Data selection algorithms can be applied as a technique to obtain data of a particular domain. Generally speaking, these methods start from a large set of sentences, and from this set select a subset of sentences that are closer to the domain of interest than other sentences in the large set. Among these methods, Transductive Algorithms (TA) perform the selection by using the test set as seed and retrieving those sentences that are relatively closer to this seed than others. Models built using the output of TA also perform better than general-domain models [1,2].

A. Gelbukh (Ed.): CICLing 2019, LNCS 13451, pp. 567–579, 2023.
https://doi.org/10.1007/978-3-031-24337-0_40

Alternatively, a general-domain model can also be adapted to a certain domain by applying the technique known as *fine-tuning* [3–5]. This consists of training the last epochs of an NMT model (built with out-domain data) using a smaller but in-domain set of sentences.

Unfortunately, additional data that are closer to the test set are not always available. The work of [6] showed that the inclusion of back-translated data can boost the performance of NMT models. Since then, adding synthetic data for training Machine Translation (MT) models has become more popular.

In this work we want to investigate whether it is useful to apply TA to synthetic data selection, in order to retrieve artificial sentences closer to the test set. We study the performance of TA on the task of synthetic data selection, applied in two different configurations (see Fig. 1):

1. Batch processing: The first approach involves back-translating a monolingual set of sentences completely and then selecting sentences from synthetic parallel set. The selection criteria of TA are based on the overlap of n-grams of the test set (the seed) with those in the source-side of the parallel set. For this reason, the performance of TA may be worse on back-translated data as the n-grams, which have been artificially generated, may be unnatural in terms of word-order.
2. Online processing: This involves selecting the necessary monolingual, target-side, sentences and afterwards back-translating the selected set. The advantage of the online process is that it is not necessary to back-translate the complete data set before selecting data. Nevertheless, as the selection is performed in monolingual target-language we cannot use the test set (which is in

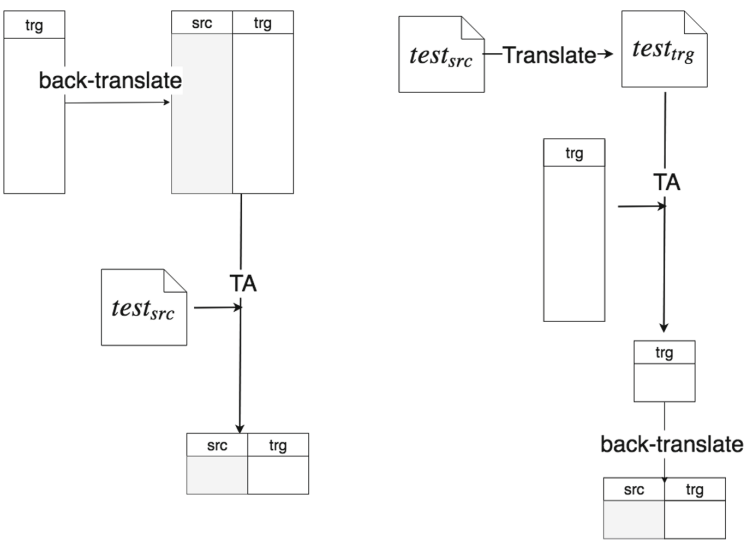

Fig. 1. Pipeline of the batch (left) and online (right) processing to obtain TA-selected synthetic data.

the source-side language) as seed. To solve this, we can proceed as described in the work of [7] and translate the test set using a generic-domain NMT model. Then, this translated text can be used as seed.

2 Related Work

2.1 Transductive Data Selection Algorithms

In this section we describe the algorithms used in the paper, which belong to the family of transductive [8] data selection methods. Such methods select the most relevant sentences for the test set using the (source-side) test set itself. The methods score each sentence s in the candidate data U (the set of sentences that have not been yet selected), and then the sentence with the highest score is added to selected pool L, which is initially empty. Note that this process is done iteratively as the scores (which depend on U and L) are updated after a sentence has been selected.

Infrequent n-gram Recovery (INR): In the work of [9,10] they propose extracting sentences containing n-grams (present in the test set) that are considered infrequent. Therefore, words such as stop words are ignored. The sentences in the candidate data U are scored according to Eq. (1):

$$score(s, U) = \sum_{ngr \in S_{test}} max(0, t - C_{S_I+L}(ngr)) \tag{1}$$

where t is the threshold that indicates whether an n-gram is frequent or not. If the count of the n-gram ngr ($C_{S_I+L}(ngr)$) in the selected pool L (and an in-domain set S_I used for initialization) exceeds the value of t then it will not contribute to the score of the sentence.

Feature Decay Algorithms (FDA): Feature Decay Algorithms [11] selects data by promoting sentences containing many n-grams from the test set, but penalizing those n-grams that have been selected several times. Each n-gram ngr is assigned an initial score, then each time a sentence containing ngr is selected the score of ngr is decreased. The default scoring function is defined as in Eq. (2):

$$score(s, L) = \frac{\sum_{ngr \in S_{test}} 0.5^{C_L(ngr)}}{length(s)} \tag{2}$$

Observe that the more occurrences of ngr are in the selected pool L ($C_L(ngr)$) the less it contributes towards the scoring of the sentence s.

2.2 Using Approximated Target Side

The methods presented in Sect. 2.1 use the test set as seed in order to retrieve sentences. However, a similar approach can be executed by using an approximated translation of the test set (approximated target side) as seed [7]. This seed can be generated by another MT model.

The output of a TA, such as INR or FDA, can be represented as a sequence of sentences $TA_{src} = (s_1^{(src)}, s_2^{(src)}, s_3^{(src)}, ...s_N^{(src)})$ of N sentences. We use the subscript *src* to indicate that the seed is a text in the source language. However, we can first translate the test set using a generic NMT model and execute the TA using the translation as a seed. The output of this execution could also be represented as a sequence of sentences $TA_{trg} = (s_1^{(trg)}, s_2^{(trg)}, s_3^{(trg)}, ...s_N^{(trg)})$

The two outputs, TA_{src} and TA_{trg}, can be combined as a new sequence of N sentences as in Eq. (3)

$$TA = (s_1^{(src)}, ...s_{N*\alpha}^{(src)}, s_1^{(trg)}, ...s_{N*(1-\alpha)}^{(trg)}) \tag{3}$$

where the top sentences from each output are concatenated. The value of $\alpha \in [0, 1]$ represents the proportion of data that are selected from TA_{src} and TA_{trg}.

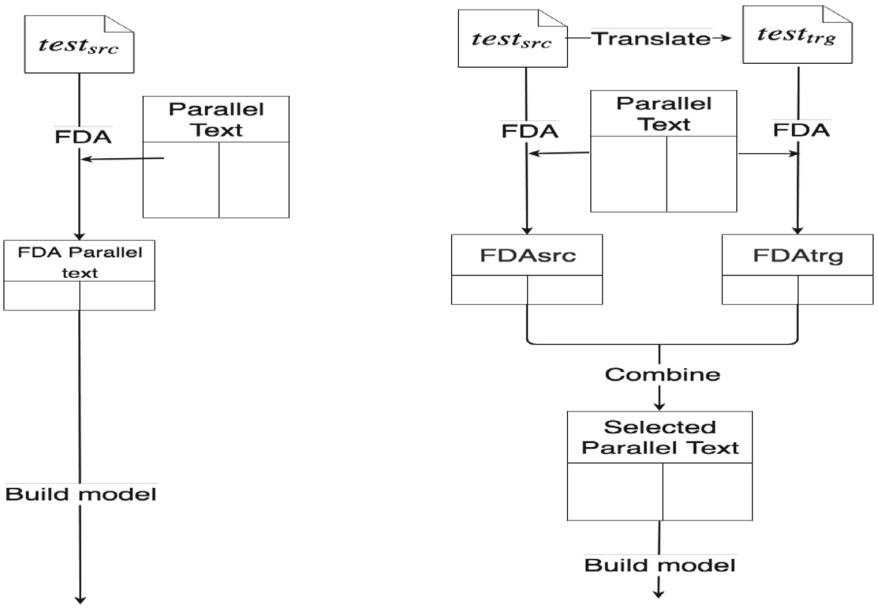

Fig. 2. Pipeline of the traditional usage of FDA (left) and pipeline of our proposal, using the target-side (right) [7].

Figure 2 (right) shows the pipeline that we followed to build the mixture of the outputs using both seeds. Although the data obtained from TA_{trg} are not always useful for adapting an MT model for the test set, mixing the data selected using the test set and the approximated target side can lead to improvements [7].

3 Fine-Tuning Models with Synthetic Data

The work of [6] showed that NMT models can be improved by adding synthetic training data. In their work they use monolingual sentences in the target language and translate them into the source-language with an NMT model. This creates a parallel corpus in which the source side has been artificially generated and the target side is human-produced data (and hence, the fluency of the translation will not be affected). Models built with back-translated data alone (or mixed with back-translated data) can have a performance comparable to those built with real data [12].

In this work we want to explore the performance of NMT models when fine-tuned with TA-selected synthetic data so they are adapted to a given test set. We are interested in exploring three main Research Questions (RQ):

- RQ1: **Does a model adapted with TA-selected back-translated data achieve improvements over the non-adapted model?**
  The strength of performing the fine-tuning technique is to adapt a model with data in the same domain as the document to be translated. Although TA can retrieve relevant data, we do not know the performance when executed using synthetic data. The artificially-generated sentences may contain unusual n-grams, so the overlap with the test set is lower. This prevents TA from retrieving relevant sentences.
- RQ2: **Does a model adapted with TA-selected back-translated data perform better than a model adapted with TA-selected authentic data?**
  Suppose that using synthetic data for adaptation leads to improvements, we also want to compare the performance to that of a model adapted with TA-retrieved authentic data. The quality of the back-translated (source) data, in terms of being an exact translation of the target, is expected to be lower than that of the source-side in the corresponding authentic sentence pairs (which were after all created by human translators). However, the authentic data have already been used to build the model to be adapted, whereas the selected artificial (source) sentences is a set of newly generated data, which may add useful new information not present in the original authentic data set. For this reason, the selected synthetic data might add more value to training the model and may also improve generalization. Therefore, fine-tuning with selected back-translated data may yield larger performance gains than fine-tuning with (repeated) authentic sentences.
- RQ3: **Is it preferable to follow the batch or the online processing?**
  As both processing (batch and online) retrieve different subsets of data, we want to study the performance of the models when they are adapted with a mixture of both outputs. The strategy we follow to combine the outputs is to concatenate them in different proportion in a similar way (using different sizes of α) as explained in Sect. 2.2.

4 Experiments

4.1 Experimental Settings

We build German-to-English models with the parallel data provided in the WMT 2015 [13] (*training data*). All data sets are tokenized and truecased. We also apply Byte Pair Encoding (BPE) [14] with 89500 merge operations. The synthetic data are built by translating the target-side (English) into the source language (German). We use an NMT model built with 1M randomly-selected sentences.

The NMT models are built using OpenNMT-py[1] [15] with the default parameter values: 2-layer LSTM with 500 hidden units, vocabulary size of 50000 words for each language.

All the models built are evaluated on two test sets using BLEU [16], TER [17] and METEOR [18] evaluation metrics. These metrics provide an estimation of the quality of the translation compared to a human-translated reference. The two test sets used to evaluate the models are: (i) *NEWS test set* provided in WMT 2015 News Translation Task; and (ii) *BIO test set*, the Cochrane[2] dataset from the WMT 2017 biomedical translation shared task [19].

In each table, we mark in bold the scores that are better than the baseline, and if they constitute a statistically significant improvement (at level $p = 0.01$) we mark them with an asterisk. This was computed with multeval [20] using bootstrap resampling [21].

4.2 Model Adaptation with Subsets of Data

The general-domain model used in this work as baseline is an NMT model trained with the complete training dataset for 13 epochs. The result of the model can be seen in Table 1

Table 1. Results of the general-domain model evaluated in the NEWS test set and BIO test set.

	NEWS	BIO
BLEU	0.2634	0.3314
TER	0.5441	0.4679
METEOR	0.3009	0.3457

The experiments carried out consist of using INR and FDA to select different sizes of data: 100 K, 200 K and 500 K sentence pairs. In INR method, a low value of t causes the method to be more strict and retrieve less sentences. We use the larger value so the execution does not exceed 48 h (i.e. $t = 80$ for NEWS test

[1] https://github.com/OpenNMT/OpenNMT-py.
[2] http://www.himl.eu/test-sets.

Table 2. Results of the models built with different sizes of INR_{src} and INR_{trg} using authentic data.

		Baseline	$\alpha = 1$	$\alpha = 0.75$	$\alpha = 0.50$	$\alpha = 0.25$	$\alpha = 0$
NEWS							
100 K	BLEU	0.2634	**0.2649**	**0.2659**	**0.2664***	**0.2655**	**0.2659***
	TER	0.5441	**0.5419**	**0.5408***	**0.5417***	**0.5413**	**0.5430***
	METEOR	0.3009	**0.3021***	**0.3030***	**0.3037***	**0.3033***	**0.3034***
200 K	BLEU	0.2634	**0.2644**	**0.2661***	**0.2666***	**0.2655**	**0.2649**
	TER	0.5441	**0.5435**	**0.5410***	**0.5406***	**0.5413***	**0.5437***
	METEOR	0.3009	**0.3012**	**0.3025***	**0.3028***	**0.3029***	**0.3027***
BIO							
100 K	BLEU	0.3314	**0.3352***	**0.3346**	**0.3347**	**0.3370***	**0.3339**
	TER	0.4679	**0.4592***	**0.4631**	**0.462**	**0.4591***	**0.4605***
	METEOR	0.3457	**0.3477**	**0.3478**	**0.3463**	**0.3488***	**0.3475**
200 K	BLEU	0.3314	**0.3388***	**0.3362***	**0.3403***	**0.3386***	**0.3343**
	TER	0.4679	**0.459***	**0.4589***	**0.457***	**0.4563***	**0.4590***
	METEOR	0.3457	**0.3494***	**0.3477**	**0.3502***	**0.3489***	**0.3495***

set and $t = 640$ for BIO test set). However, the amount of sentences retrieved are below 500 K, so in the experiments we only evaluate the models adapted with 100 K and 200 K INR-selected sentences. The sentences retrieved are used to adapt the general-domain model. In particular, we adapt the 12th epoch of the model by fine-tuning it with the selected data.

In Table 2 and Table 3 we show the performance of the models when fine-tuned with different sizes of selected authentic data. In the tables we also indicate the proportions of data selected using the test set or the approximated target side as seed.

As we can see, the performance of the adapted models are higher than that of the general-domain model (Table 1). In addition, using a mixture of TA_{src} and TA_{trg} (columns $\alpha = 0.75$, $\alpha = 0.50$ and $\alpha = 0.25$) can achieve a higher performance than TA_{src} or TA_{trg} alone.

In our experiments we follow the same procedure using synthetic data in order to perform comparisons among the general-domain model, models adapted with authentic data, and models adapted with synthetic data.

5 Results

The results of the models adapted with synthetic data are shown in Table 4 (INR method) and Table 5 (FDA method). In order to answer RQ1, we include in the first column, as baseline, the performance of the 13th epoch of the general-domain model (Table 1). We mark in bold those scores that indicate a better

Table 3. Results of the models built with different sizes of FDA_{src} and FDA_{trg} using authentic data.

		Baseline	$\alpha = 1$	$\alpha = 0.75$	$\alpha = 0.50$	$\alpha = 0.25$	$\alpha = 0$
NEWS							
100 K	BLEU	0.2634	**0.2649**	**0.2665***	**0.2642***	**0.2643**	0.2633
	TER	0.5441	**0.5421**	**0.5412***	**0.5413***	**0.5416***	**0.5416***
	METEOR	0.3009	**0.3021***	**0.3027***	**0.3022***	**0.3019**	**0.3020**
200 K	BLEU	0.2634	**0.2655**	**0.2665***	**0.2651**	**0.2652**	**0.2654***
	TER	0.5441	**0.5417***	**0.5412***	**0.5413***	**0.5421***	**0.5404***
	METEOR	0.3009	**0.3024***	**0.3027***	**0.3025***	**0.3025***	**0.3027***
500 K	BLEU	0.2634	**0.264***	**0.2658***	**0.2671***	**0.2654**	**0.2650**
	TER	0.5441	0.5447	**0.5414***	**0.5412***	**0.5415***	**0.5404***
	METEOR	0.3009	**0.3010***	**0.3028***	**0.3028***	**0.3024***	**0.3028***
BIO							
100 K	BLEU	0.3314	**0.3368***	**0.3377***	**0.3391***	**0.339***	**0.3331**
	TER	0.4679	**0.4597***	**0.4611***	**0.4599***	**0.4597***	**0.4649**
	METEOR	0.3457	**0.3471**	**0.3473**	**0.3476**	**0.3485**	**0.3463**
200 K	BLEU	0.3314	**0.3396***	0.3414*	**0.3375***	**0.3391***	**0.3370***
	TER	0.4679	**0.4564***	**0.459***	**0.4574***	**0.4596***	**0.4572***
	METEOR	0.3457	**0.3501***	**0.3503***	**0.3491***	**0.3484***	**0.3496***
500 K	BLEU	0.3314	**0.3375***	**0.3406***	**0.3358***	**0.3354***	**0.3336**
	TER	0.4679	**0.4592***	**0.4552***	**0.4593***	**0.4574***	**0.4617**
	METEOR	0.3457	**0.3492***	**0.3496***	**0.3485**	**0.3494***	**0.3485***

performance than the baseline and add an asterisk if they are statistically significant at level $p = 0.01$.

In the tables we observe that adapted models with artificial data tend to perform better on NEWS test set than BIO test set (e.g. BLEU scores are only higher in the NEWS test set). This manifests that the domain of the model used for back-translating plays an important role. In our experiments the above model is closer to the news domain because it was built using a sample of the authentic training data.

METEOR scores of adapted models are higher than those of the general-domain model for both test sets, and in many cases the improvements are statistical significant (with $p = 0.001$). In contrast, TER scores are lower than the baseline. This may be caused by the synonym or conjugation chosen by the adapted model. For example, the sentence "auch Schulen" is translated by the general-domain model as "schools too" (the same as in the reference), but adapted model produced "also schools".

Table 4. Results of the models built with different sizes of INR_{src} and INR_{trg} using back-translated data.

		Baseline	$\alpha = 1$	$\alpha = 0.75$	$\alpha = 0.50$	$\alpha = 0.25$	$\alpha = 0$
NEWS							
100 K	BLEU	0.2634	**0.2664**	**0.267**	**0.2671**	**0.2679***	**0.2675***
	TER	0.5441	0.5492	0.5496	0.55	0.5496	0.5513
	METEOR	0.3009	**0.3058***	**0.3062***	**0.3063***	**0.3067***	**0.3061***
200 K	BLEU	0.2634	**0.2666**	**0.2673***	**0.2678***	**0.2673***	**0.2672***
	TER	0.5441	0.5485	0.5486	0.5478	0.5481	0.5481
	METEOR	0.3009	**0.3064***	**0.3061***	**0.3068***	**0.3066***	**0.3068***
BIO							
100 K	BLEU	0.3314	0.324	0.327	0.3263	0.3269	0.3251
	TER	0.4679	0.4762	0.4747	0.4753	0.4751	0.4764
	METEOR	0.3457	**0.3486**	**0.3490**	**0.3502***	**0.351***	**0.3489**
200 K	BLEU	0.3314	0.3241	0.3255	0.3255	0.3254	0.3251
	TER	0.4679	0.4782	0.4755	0.4732	0.4742	0.4745
	METEOR	0.3457	**0.3487**	**0.3501***	**0.3508***	**0.3509***	**0.3505***

5.1 Model Adaptation with Synthetic Data

In our experiments, the back-translated data used for the adaptation are new data unseen by the model (the authentic data used to adapt the models presented in Tables 2 and 3 are subsets of the same data used to build the general-domain model). The outcomes observed in the experiments show that adapting the models with synthetic data does not achieve as good results as adapting them with authentic data (which answers the RQ2). If we compare cell-wise (i.e. same value of α and same size of selected sentences) Tables 2 and 4 or Tables 3 and 5 we see slight improvements for the BLEU and METEOR scores for the news test set (NEWS subtables). However, none of these are statistically significant at $p = 0.01$.

As mentioned previously, the sentences produced by the model used for back-translation may contain mistakes such as word-ordering, incorrect translations etc. which reduces the potential sentences that TA can retrieve. For example, in our experiments we find the following sentence in the NEWS test set "Auf der Hüpfburg beim Burggartenfest war am Sonnabend einiges los." (according to the reference "Something is happening on the bouncy castle at the Burggartenfest.") contains the word "Hüpfburg" ("bouncy castle") which is used by TA to retrieve sentences. There are 18 occurrences of this word in the authentic data set. However, in the synthetic data there are no instances of this word. Instead, the back-translated counterparts of sentences containing "Hüpfburg" include words such as "bouncer" (copied from the English side) or "bounmit" (a word that does not exist). Nevertheless, in some cases back-translated sentences may be closer to literal translation than those found in the authentic

Table 5. Results of the models built with different sizes of FDA_{src} and FDA_{trg} using back-translated data.

		Baseline	$\alpha = 1$	$\alpha = 0.75$	$\alpha = 0.50$	$\alpha = 0.25$	$\alpha = 0$
NEWS							
100 K	BLEU	0.2634	**0.2639**	**0.2654**	**0.264**	**0.2655**	**0.2672***
	TER	0.5441	0.5525	0.5509	0.5522	0.5511	0.5493
	METEOR	0.3009	**0.305***	**0.3054***	**0.3051***	**0.3055***	**0.3062***
200 K	BLEU	0.2634	**0.2655**	**0.2658**	**0.2663**	**0.2666**	**0.2679***
	TER	0.5441	0.5497	0.5512	0.5504	0.5493	0.5484
	METEOR	0.3009	**0.3051***	**0.3053***	**0.306***	**0.3055***	**0.3063***
500 K	BLEU	0.2634	**0.2662**	**0.2674***	**0.2668**	**0.2679***	**0.2664**
	TER	0.5441	0.5483	0.5494	0.5501	0.5488	0.5489
	METEOR	0.3009	**0.3061***	**0.3068***	**0.3062***	**0.3068***	**0.3062***
BIO							
100 K	BLEU	0.3314	0.3228	0.3248	0.3238	0.3254	0.3262
	TER	0.4679	0.4755	0.475	0.4751	0.4742	0.4744
	METEOR	0.3457	**0.349**	**0.3488**	**0.3497***	**0.3521***	**0.3500***
200 K	BLEU	0.3314	0.3214	0.3245	0.3258	0.3255	0.3241
	TER	0.4679	0.478	0.4743	0.4737	0.4751	0.4749
	METEOR	0.3457	**0.3487**	**0.3495**	**0.3501***	**0.349**	**0.3482**
500 K	BLEU	0.3314	0.3215	0.3223	0.3229	0.3241	0.3226
	TER	0.4679	0.4842	0.4843	0.4817	0.4813	0.4811
	METEOR	0.3457	**0.3478**	**0.3488**	**0.3486**	**0.3491**	**0.349**

set [7,22]. For example, in the authentic data set we find the sentence-pair ⟨"er ist verheiratet und hat zwei Kinder.","since then, he has had a long career on stage, in film and on television. He has also established himself as a singer and an author in recent years."⟩ which do not convey the same meaning. However, the machine-produced source-side is "seitdem hat er eine lange Karriere auf der Bühne, im Film und im Fernsehen absolviert und hat sich auch als Sängerin und Autor in den letzten Jahren etabliert" which is closer in meaning to the target-side sentence. Another example is the pair ⟨"10 %!","one tenth!"⟩. Although, they have the same meaning, in the back-translated counterpart the source-side sentence is "ein Zehntel!", which is a literal translation.

5.2 Batch and Online Processing

In order to answer RQ3 we need to compare columns $\alpha = 1$ (batch processing, i.e. extract from back-translated data using the test set) and $\alpha = 0$ (online processing, i.e. extract from authentic data using the approximated set). In Table 4 and Table 5 we see that in our experiments following the online process the results tend to be better.

Using an approximated target side as seed is risky, as it can be of low quality. For example, the sentence "Das Buch wurde neu für 48$ verkauft." ("The book was selling for $48 new.") is translated as "The book was sold for 48$." by the general-domain model. As we can see, the word "new" is omitted in the translation. This means that the TA will not consider the word "new" when selecting sentences.

Despite that, we find that the generated target-side seed may contain n-grams that better represent the context of the input document. For example, the sentence in the test set "Ich liebe es, in einem Probenraum zu sein." is translated, according to the reference, as "I love being in a rehearsal room.". The model adapted with 100 K sentences from FDA_{src} ($\alpha = 1$) generates the translation "I love to be in a sample room.", whereas the model adapted with FDA_{trg} ($\alpha = 0$) produces a sentence that conveys the same meaning to the reference: "I love to be in a rehearsal room.".

We observe that the occurrences of "Proben" (due to BPE, the word is splitted as "Proben@@ raum") are translated as "sample" or "rehersal" depending on the context. The fact that in the approximated target side the word has been accurately translated as "rehearsal room" induces FDA_{trg} to select more sentences that include the term "rehearsal". In contrast, FDA_{src} retrieves sentences based on the word "Proben" in the seed (as it is present in the test set). However, in the training data this word has been artificially produced and it replaces words such as "Messwasser" ("water sample") or "Musterproduktion" ("sample production").

6 Conclusion and Future Work

In this paper we have analyzed various use-cases of synthetic data for adapting a general-domain model. We have seen that using a TA it is possible to obtain sentences from synthetic data that can improve the model, even if the sentences used for adaptation are an artificial version of the same sentences used to construct the general model.

In addition, we have seen that performing the adaptation online, extracting just the necessary monolingual target-language sentences (using an approximated translation of the test set as seed) and back-translating them afterwards, is a reasonable approach that can even perform better than selecting directly from synthetic sentences.

In the future, we want to further extend this research and explore the effects on the performance of combining both authentic and synthetic data or the use of forward-translation [23]. In addition, we are interested in exploring whether the results observed in this paper are the same when using other language pairs or other configurations of INR and FDA [24,25].

Acknowledgements. This research has been supported by the ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

 This work has also received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 713567.

References

1. Poncelas, A., de Buy Wenniger, G.M., Way, A.: Feature decay algorithms for neural machine translation. In: Proceedings of the 21st Annual Conference of the European Association for Machine Translation, pp. 239–248. Alacant, Spain (2018)
2. Silva, C.C., Liu, C.H., Poncelas, A., Way, A.: Extracting in-domain training corpora for neural machine translation using data selection methods. In: Proceedings of the Third Conference on Machine Translation: Research Papers, pp. 224–231, Brussels, Belgium (2018)
3. Luong, M.T., Manning, C.D.: Stanford neural machine translation systems for spoken language domains. In: Proceedings of the International Workshop on Spoken Language Translation, pp. 76–79. Da Nang, Vietnam (2015)
4. Freitag, M., Al-Onaizan, Y.: Fast domain adaptation for neural machine translation. arXiv preprint arXiv:1612.06897 (2016)
5. van der Wees, M., Bisazza, A., Monz, C.: Dynamic data selection for neural machine translation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1400–1410. Copenhagen, Denmark (2017)
6. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 86–96. Berlin, Germany (2016)
7. Poncelas, A., de Buy Wenniger, G.M., Way, A.: Data selection with feature decay algorithms using an approximated target side. In: 15th International Workshop on Spoken Language Translation, pp. 173–180. Bruges, Belgium (2018)
8. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
9. Parcheta, Z., Sanchis-Trilles, G., Casacuberta, F.: Data selection for NMT using infrequent n-gram recovery. In: Proceedings of the 21st Annual Conference of the European Association for Machine Translation, pp. 219–227. Alacant, Spain (2018)
10. Gascó, G., Rocha, M.A., Sanchis-Trilles, G., Andrés-Ferrer, J., Casacuberta, F.: Does more data always yield better translations? In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 152–161. Avignon, France (2012)
11. Biçici, E., Yuret, D.: Instance selection for machine translation using feature decay algorithms. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, pp. 272–283. Edinburgh, Scotland (2011)
12. Poncelas, A., Shterionov, D., Way, A., de Buy Wenniger, G.M., Passban, P.: Investigating backtranslation in neural machine translation. In: 21st Annual Conference of the European Association for Machine Translation, pp. 249–258. Alacant, Spain (2018)
13. Bojar, O., et al.: Findings of the 2015 workshop on statistical machine translation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation, pp. 1–46. Lisboa, Portugal (2015)
14. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1715–1725. Berlin, Germany (2016)

15. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.M.: OpenNMT: open-source toolkit for neural machine translation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pp. 67–72. Vancouver, Canada (2017)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. Philadelphia, Pennsylvania, USA (2002)
17. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pp. 223–231. Cambridge, Massachusetts, USA (2006)
18. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72. Ann Arbor, Michigan (2005)
19. Yepes, A.J., et al.: Findings of the WMT 2017 biomedical translation shared task. In: Proceedings of the Second Conference on Machine Translation, pp. 234–247 (2017)
20. Clark, J.H., Dyer, C., Lavie, A., Smith, N.A.: Better hypothesis testing for statistical machine translation: controlling for optimizer instability. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 2, pp. 176–181. Portland, Oregon (2011)
21. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 388–395. Barcelona, Spain (2004)
22. Poncelas, A., Way, A., Sarasola, K.: The ADAPT system description for the IWSLT 2018 Basque to English translation task. In: 15th International Workshop on Spoken Language Translation, pp. 76–82. Bruges, Belgium (2018)
23. Chinea-Rios, M., Peris, A., Casacuberta, F.: Adapting neural machine translation with parallel synthetic data. In: Proceedings of the Second Conference on Machine Translation, pp. 138–147. Copenhagen, Denmark (2017)
24. Poncelas, A., Way, A., Toral, A.: Extending feature decay algorithms using alignment entropy. In: Quesada, J.F., Martín Mateos, F.J., López-Soto, T. (eds.) FETLT 2016. LNCS (LNAI), vol. 10341, pp. 170–182. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69365-1_14
25. Poncelas, A., de Buy Wenniger, G.M., Way, A.: Applying n-gram alignment entropy to improve feature decay algorithms. Prague Bull. Math. Linguist. **108**, 245–256 (2017)