# Selecting Artificially-Generated Sentences for Fine-Tuning Neural Machine Translation

**Alberto Poncelas** and **Andy Way**
ADAPT Centre, School of Computing,
Dublin City University, Dublin, Ireland
{firstname.lastname}@adaptcentre.ie

## Abstract

Neural Machine Translation (NMT) models tend to achieve best performance when larger sets of parallel sentences are provided for training. For this reason, augmenting the training set with artificially-generated sentence pairs can boost performance.

Nonetheless, the performance can also be improved with a small number of sentences if they are in the same domain as the test set. Accordingly, we want to explore the use of artificially-generated sentences along with data-selection algorithms to improve German-to-English NMT models trained solely with authentic data.

In this work, we show how artificially-generated sentences can be more beneficial than authentic pairs, and demonstrate their advantages when used in combination with data-selection algorithms.

## 1 Introduction

The data used for training Machine Translation (MT) models consist mainly of a set of parallel sentences (a set of sentence-pairs in which each sentence is paired with its translation). As Neural Machine Translation (NMT) models typically achieve best performance when using large sets of parallel sentences, they can benefit from the sentences created by Natural Language Generation (NLG) systems. Although artificial data is expected to be of lower quality than authentic sentences, it still can help the model to learn how to better generalize over the training instances and produce better translations.

A popular technique used to create artificial data is the back-translation technique (Sennrich et al., 2016a; Poncelas et al., 2018c). This consists of generating sentences in the source language by translating monolingual sentences in the target language. Then, these sentences in both languages are paired and can be used to augment the original parallel training set used to build better NMT models.

Nonetheless, if synthetic data are not in the same domain as the test set, it can also hurt the performance. For this reason, we explore an alternative approach to better use the artificially-generated training instances to improve NMT models. In particular, we propose that instead of blindly adding back-translated sentences into the training set they can be considered as candidate sentences for a data-selection algorithm to decide which sentence-pairs should be used to fine-tune the NMT model. By doing that, instead of increasing the number of training instances in a motivated manner, the generated sentences provide us with more chances of obtaining relevant parallel sentences (and still use smaller sets for fine-tuning).

As we want to build task-specific NMT models, in this work we explore two data-selection algorithms that are classified as Transductive Algorithms (TA): Infrequent *N*-gram Recovery (INR) and Feature Decay Algorithms (FDA). These methods use the test set $S_{test}$ (the document to be translated) as the seed to retrieve sentences. In transductive learning (Vapnik, 1998) the goal is to identify the best training instances to learn how to classify a given test set. In order to select these sentences, the TAs search for those *n*-grams in the test set that are also present in the source side of the candidate sentences.

Although augmenting the candidate pool with more sentences should be beneficial, as the TAs select the sentences based on overlapping *n*-gram the mistakes produced by the model used for back-translation (which are those commonly addressed in NLG such as the generated word order or word choice) can be a disadvantage.

In this work, we explore whether TAs are more inclined to select authentic or artificial sentences. In addition, we propose three different methods of

how they can be combined into a single hybrid set. Finally, we investigate whether the hybrid sets retrieved by TAs can be more useful than the authentic set of sentences to fine-tune NMT models.

## 2 Related Work

The work presented in this paper is based on two main concepts: the generation of synthetic sentences, and the selection of sentences from a set $S$ of candidates.

### 2.1 Use of Artificially-Generated Data to Improve MT Models

The proposal of Sennrich et al. (2016a) showed that NMT models can be improved by back-translating a set of (monolingual) sentences in the target side into the source side using an MT model. Other uses of monolingual target-side sentences include building the parallel set by using a NULL token in the source side (Sennrich et al., 2016a) or creating language models to improve the decoder (Gülçehre et al., 2015).

Hoang et al. (2018) improve the model used for back-translation by training this model with increasing amounts of artificial sentences. They iteratively improve the models creating artificial sentences of better quality.

Similarly to this paper, the use of artificially-generated sentences to fine-tuned models has also been explored by Chinea-Rios et al. (2017) where they select monolingual authentic sentences in the source-side and translate them into the target language, or the work of Poncelas et al. (2019a) where they use back-translated sentences only to adapt the models.

### 2.2 Adaptation of NMT Models to the Test Set

The improvement of NMT models can be performed by fine-tuning (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016), i.e. train the models for additional epochs using a small set of in-domain data. Alternatively, van der Wees et al. (2017) train models using smaller but more in-domain sentences in each epoch of the training process.

The use of the test set to retrieve relevant sentences for fine-tuning the model has been explored by Li et al. (2016), adapting a different model for each sentence in the test set, or Poncelas et al. (2018b, 2019b) where they adapt the model for the complete test set using transductive data-selection algorithms.

## 3 Transductive Algorithms

In this paper, the sentences used to fine-tune the model are retrieved using INR and FDA. These methods select sentences by scoring each sentence $s$ from the candidate pool $S$, and adding that with the highest score to a selected pool $L$. This process is performed iteratively until the selected pool contains $N$ sentences.

**Infrequent *n*-gram Recovery** (INR) (Parcheta et al., 2018; Gascó et al., 2012): This method selects those sentences that contain *n*-grams from the test set that are infrequent (ignoring frequent words such as stop words or general-domain terms). A candidate sentence $s \in S$ is scored according to the number of infrequent *n*-grams shared with the set of sentences of the test set $S_{test}$, computed as in (1):

$$score(s) = \sum_{ngr \in \{S_{test} \bigcap s\}} max(0, t - C_L(ngr)) \quad (1)$$

where $t$ is the threshold that indicates the number of occurrences of an *n*-gram to be considered infrequent. If the number of occurrences of $ngr$ in the selected pool ($C_L(ngr)$) is above the threshold $t$, then the component $max(0, t - C_S(ngr))$ is 0 and so the *n*-gram does not contribute to scoring the sentence.

**Feature Decay Algorithms** (Biçici and Yuret, 2011; Biçici, 2013) also retrieve those sentences sharing the highest number of *n*-grams from the test set. However, in order to increase the variability and avoid selecting the same *n*-grams, those that have been selected are penalized is proportional to the number of occurrences in $L$. The score of a sentence is computed as in (2):

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

where $length(s)$ indicates the number of words in the sentence $s$. According to the equation, the more occurrences of $ngr$ in $L$, the smaller the contribution is to the scoring of the sentence $s$.

### 3.1 Models Adapted with Hybrid Data

In order to fine-tune models with hybrid data, we propose three methods of creating these sets: *hybr*, *batch* and *online*. These methods can be classified depending on whether the combination is performed before or after the execution of the TA.

**Combine Before Selection.** This approach consists of selecting from a hybrid set (*hybr*). This involves concatenating both the authentic candidate $S_{auth}$ and artificial $S_{synth}$ sentences as a first step and then executing the TAs with the new candidate set $S_{auth+synth}$.

**Combine After Selection.** Another approach is to force the presence of both authentic and synthetic sentences by using different proportions of TA-selected authentic ($L_{auth}$) and synthetic ($L_{synth}$) sentence pairs. We concatenate the top-$(N * \gamma)$ sentences from the selected authentic set and the top-$(N*(1-\gamma))$ from the synthetic set. The value of $\gamma \in [0, 1]$ indicates the proportion of authentic and synthetic sentences. For example, $\gamma = 0.75$ indicates that the 75% of sentences in the dataset are authentic and the remaining 25% are artificially generated.

The selected synthetic set $L_{synth}$ can be obtained by executing the TAs on artificial candidate sentences $S_{synth}$ (*batch*). This implies that the sentences will be retrieved by finding overlaps of *n*-grams between the test set and artificial sentences.

Alternatively, the retrieval may be carried by finding overlaps in the target-side (*online*) as they are human-produced sentences. However, as the test set is in the source language, we need to first generate an approximated translation of the test with a general-domain MT model (Poncelas et al., 2018a,d). Unlike in *batch*, the advantage of this approach is that it is not necessary to generate the source side of the whole set of monolingual sentences, but rather only those selected by the TA.

## 4 Experiments

### 4.1 Data and Models Settings

We build German-to-English NMT models using the following datasets:

- Training data: German-English parallel sentences provided in WMT 2015 (Bojar et al., 2015) (4.5M sentence pairs).
- Test sets: We evaluate the models with two test sets in different domains:
  - *BIO test set*: the Cochrane[1] dataset from the WMT 2017 biomedical translation shared task (Yepes et al., 2017).
  - *NEWS test set*: The test set provided in WMT 2015 News Translation Task.

All these data sets are tokenized, truecased, and Byte Pair Encoding (BPE) (Sennrich et al., 2016b) is applied using 89,500 merge operations.

The NMT models are built using the attentional encoder-decoder framework with OpenNMT-py[2] (Klein et al., 2017). We use the default values in the parameters: 2-layer LSTM (Hochreiter and Schmidhuber, 1997) with 500 hidden units. The size of the vocabulary is 50,000 words for each language.

In order to retrieve sentences, we use the TAs with default configuration (using *n*-grams of order 3 to find overlaps between the seed and the training data) to extract sets of 100K, 200K, and 500K sentences. We use a threshold of $t = 40$ for INR although this causes the INR to retrieve less than 500K sentences. Accordingly, the results shown for INR will include only 100K and 200K sentences.

### 4.2 Back-Translation Generation Settings

In order to generate artificial sentences, we use an NMT model (we refer to it as *BT model*) to back-translate sentences from the target language into the source language. This model is built by training a model with 1M sentences sampled from the training data and using the same configuration described above (but in the reverse language direction, English-to-German).

As we want to compare authentic and synthetic sentences, we back-translate the target-side of the training data using the BT model. By doing this we ensure both sets are comparable which allows us to perform a fair analysis of whether artificial sentences are more likely to be selected by a TA and which are more useful to fine-tune the models.

Note also that there are 1M sentences that have been generated by translating the same target-side sentences used in training. This could cause the generated sentences to be exactly the same as authentic ones. However, this is not always the case as we report in Section 5.2.

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

## 5 Results

| | | BASE13 | BASE12 + INR | BASE12 + FDA |
|---|---|---|---|---|
| BIO | | | | |
| 100K lines | BLEU | 33.14 | **33.52*** | **33.68*** |
| | TER | 46.79 | **45.92*** | **45.97*** |
| | MET. | 34.57 | **34.77** | **34.71** |
| | CHRF3 | 59.08 | **59.43** | **59.24** |
| 200K lines | BLEU | 33.14 | **33.88*** | **33.96*** |
| | TER | 46.79 | **45.90*** | **45.64*** |
| | MET. | 34.57 | **34.94*** | **35.01*** |
| | CHRF3 | 59.08 | **59.56** | **59.56** |
| 500K lines | BLEU | 33.14 | - | **33.75*** |
| | TER | 46.79 | - | **45.92*** |
| | MET. | 34.57 | - | **34.92*** |
| | CHRF3 | 59.08 | - | **59.57** |
| NEWS | | | | |
| 100K lines | BLEU | 26.34 | **26.49** | **26.49** |
| | TER | 54.41 | **54.19** | **54.21** |
| | MET. | 30.09 | **30.21*** | **30.21*** |
| | CHRF3 | 51.71 | **51.78** | **51.80** |
| 200K lines | BLEU | 26.34 | **26.44** | **26.55*** |
| | TER | 54.41 | **54.35** | **54.17*** |
| | MET. | 30.09 | **30.12** | **30.24*** |
| | CHRF3 | 51.71 | 51.67 | **51.89** |
| 500K lines | BLEU | 26.34 | - | **26.40*** |
| | TER | 54.41 | - | 54.47 |
| | MET. | 30.09 | - | **30.10*** |
| | CHRF3 | 51.71 | - | **51.71** |

Table 1: Performance of the BASE13 model, and the models fine-tuned with subsets of the training data.

First of all, we present in Table 1 the performance of the model trained with all data for 13 epochs (BASE13), as this is when the model converges. We also show the performance of the model when fine-tuning the 12th epoch with the subset of (authentic) data selected by INR (*INR* column) and FDA (*FDA* column).

In order to evaluate the performance of the models, we present the following evaluation metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2006), METEOR (Banerjee and Lavie, 2005), and CHRF (Popovic, 2015). These metrics provide an estimation of the translation quality when the output is compared to a human-translated reference. Note that in general, the higher the score, the better the translation quality is. The only exception is TER which is an error metric and so lower results indicate better quality.

In addition, we indicate in bold those scores that show an improvement over the baseline (in Table 1 we use BASE13 as the baseline) and add an asterisk if the improvements are statistically significant at p=0.01 (using Bootstrap Resampling (Koehn, 2004), computed with multeval (Clark et al., 2011)).

In the table, we can see that using a small subset of data for training the 13th epoch can cause the performance of the model to improve. In the following experiments, we want to compare whether augmenting the candidate set with synthetic data can further boost these improvements. For this reason, we use *INR* and FDA *FDA* as baselines.

### 5.1 Results of Models Fine-tuned with Hybrid Data

| | | INR | INR HYBR |
|---|---|---|---|
| BIO | | | |
| 100K lines | BLEU | 33.52 | **33.87** |
| | TER | 45.92 | 46.17 |
| | METEOR | 34.77 | **35.01** |
| | CHRF3 | 59.43 | **59.53** |
| 200K lines | BLEU | 33.88 | 33.70 |
| | TER | 45.90 | 46.33 |
| | METEOR | 34.94 | **35.23** |
| | CHRF3 | 59.56 | **60.03** |
| NEWS | | | |
| 100K lines | BLEU | 26.49 | **26.76** |
| | TER | 54.19 | 54.36 |
| | METEOR | 30.21 | **30.48*** |
| | CHRF3 | 51.78 | **52.35*** |
| 200K lines | BLEU | 26.44 | **26.80*** |
| | TER | 54.35 | **54.34** |
| | METEOR | 30.12 | **30.51*** |
| | CHRF3 | 51.67 | **52.39*** |

Table 2: Results of the models built with different sizes of INR-selected hybrid data following the *hybr* approach. The results in bold indicate an improvement over INR. The asterisk means the improvement is statistically significant at p=0.01.

In the first set of experiments we explore the *hybr* approach, i.e. the TAs are executed on a mixture of authentic and synthetic data (combined before the execution of the TA). We present the results of the models trained with these sets in Table 2 (for INR) and Table 3 (for FDA). In the first column, we include as the baseline the fine-tuned models

| | | FDA | FDA HYBR |
|---|---|---|---|
| BIO | | | |
| 100K lines | BLEU | 33.68 | **33.86** |
| | TER | 45.97 | 46.20 |
| | METEOR | 34.71 | **35.22*** |
| | CHRF3 | 59.24 | **59.89*** |
| 200K lines | BLEU | 33.96 | 33.94 |
| | TER | 45.64 | 46.09 |
| | METEOR | 35.01 | **35.29** |
| | CHRF3 | 59.56 | **59.89** |
| 500K lines | BLEU | 33.75 | **33.90** |
| | TER | 45.92 | 46.34 |
| | METEOR | 34.92 | **35.12** |
| | CHRF3 | 59.57 | **59.80** |
| NEWS | | | |
| 100K lines | BLEU | 26.49 | **26.71** |
| | TER | 54.21 | 54.31 |
| | METEOR | 30.21 | **30.43*** |
| | CHRF3 | 51.80 | **52.30*** |
| 200K lines | BLEU | 26.55 | **26.78*** |
| | TER | 54.17 | 54.42 |
| | METEOR | 30.24 | **30.51*** |
| | CHRF3 | 51.89 | **52.41*** |
| 500K lines | BLEU | 26.40 | **26.78*** |
| | TER | 54.47 | **54.38** |
| | METEOR | 30.10 | **30.51*** |
| | CHRF3 | 51.71 | **52.38*** |

Table 3: Results of the models built with different sizes of FDA-selected hybrid data following the *hybr* approach. The results in bold indicate an improvement over FDA. The asterisk means the improvement is statistically significant at p=0.01.

presented in Table 1.

The results in the tables show that increasing the size of the candidate pool is beneficial. We see that most scores are better (marked in bold) than the model fine-tuned with only authentic data. However, the performance is also dependent on the domain. When comparing BIO and NEWS subtables we see that the models adapted for the latter domain tend to achieve better performances as most of the scores are statistically significant improvements.

When analyzing the selected dataset we find that the authentic sentences constitute slightly above half (between 51% and 64% of the sentences). This is an indicator that artificially-generated sentences contain *n*-grams that can be found by TA and are as useful as authentic sentences.

In addition, the amount of duplicated target-side sentences is very low (between 10% and 13%). This indicates that the MT-generated sentences contain *n*-grams that are different from the authentic counterpart, which increases the variety of the candidates that are useful for the TA to select.

In Table 4 and Table 5 we present the results of the models when fine-tuned with a combination of authentic and synthetic data following the *Combine Before Selection* approaches in Section 3.1. The tables are structured in two subtables showing the results of *batch* and *online* approaches. Each subtable present the results of three values of $\gamma$: 0.75, 0.50 and 0.25.

In these tables, we see that the performance of the models following the *batch* and *online* approaches is similar. These results are also in accord with those obtained following the *hybr* approach, as the improvements depend more on the domain (most evaluation scores in the NEWS test set indicate statistically significant improvements whereas for the BIO test set most of them are not) than the TA used, or the value of $\gamma$. Although the best scores tend to be when $\gamma = 0.50$ this is not always the case, and moreover we can find experiments in which using high amounts of synthetic sentences (i.e. $\gamma = 0.25$) achieve better results than using a higher proportion of authentic sentences. For instance, in BIO subtable of Table 4, using 100K sentences with the *online* $\gamma = 0.25$ approach, the improvements are statistically significant for two evaluation metrics whereas in the other experiments in that row they are not.

When analyzing the translations produced by these models we find several examples in which the translations of models fine-tuned with hybrid data are superior to those tuned with authentic sentences. An example of this is the sentence in the NEWS test set "nach Krankenhausangaben wurde ein Polizist verletzt." (in the reference, "according to statements released by the hospital, a police officer was injured.")

This sentence is translated by INR and FDA models (those fine-tuned with 100K authentic sentences) as "a policeman was injured after hospital information.". We see that these models translate the word "nach" with its literal meaning ("after") whereas in this context ("nach Krankenhausangaben") it should have been translated as "according

| | | | batch | | | online | | |
|---|---|---|---|---|---|---|---|---|
| | | INR | $\gamma = 0.75$ | $\gamma = 0.50$ | $\gamma = 0.25$ | $\gamma = 0.75$ | $\gamma = 0.50$ | $\gamma = 0.25$ |
| BIO | | | | | | | | |
| 100K lines | BLEU | 33.52 | **33.88** | 33.50 | **33.67** | 33.46 | 33.62 | 33.13 |
| | TER | 45.92 | 46.23 | 46.63 | 46.60 | 46.40 | 46.39 | 46.64 |
| | METEOR | 34.77 | **35.03** | **35.13** | **35.13** | **34.90** | **34.96** | **35.12** |
| | CHRF3 | 59.43 | **59.62** | **59.72** | **59.95** | **59.48** | **59.73** | 59.92 |
| 200K lines | BLEU | 33.88 | 33.85 | 33.70 | 33.51 | 33.69 | 33.52 | 33.39 |
| | TER | 45.90 | 46.43 | 46.14 | 46.80 | 46.40 | 46.46 | 46.62 |
| | METEOR | 34.94 | **35.00** | **35.20** | **35.06** | **35.04** | **34.96** | **35.08** |
| | CHRF3 | 59.56 | 59.51 | **59.95** | **59.93** | **59.61** | 59.49 | **59.91** |
| NEWS | | | | | | | | |
| 100K lines | BLEU | 26.49 | **26.81*** | **26.59** | **26.75** | **26.77*** | **26.73** | **26.75** |
| | TER | 54.19 | 54.39 | 54.46 | 54.49 | 54.4 | 54.71 | 54.84 |
| | METEOR | 30.21 | **30.45*** | **30.51*** | **30.65*** | **30.54*** | **30.51*** | **30.64*** |
| | CHRF3 | 51.78 | **52.30*** | **52.46*** | **52.69*** | **52.34*** | **52.46*** | **52.70*** |
| 200K lines | BLEU | 26.44 | **26.85*** | **26.77*** | **26.61** | **26.79*** | **26.81*** | **26.66** |
| | TER | 54.35 | 54.37 | 54.43 | 54.70 | 54.4 | 54.67 | 54.67 |
| | METEOR | 30.12 | **30.49*** | **30.58*** | **30.60*** | **30.48*** | **30.55*** | **30.63*** |
| | CHRF3 | 51.67 | **52.35*** | **52.55*** | **52.60*** | **52.25*** | **52.57*** | **52.69*** |

Table 4: Results of the models built with different sizes of INR-selected hybrid data following the *batch* and *online* approaches. The results in bold indicate an improvement over INR. The asterisk means the improvement is statistically significant at p=0.01.

to" as stated in the reference.

In the hybrid models, we see that the same sentence has been translated as "according to hospital information, a policeman was injured." (in this case the models fine-tuned with hybrid data have produced the same translations). The models tuned with hybrid data are capable of producing the *n*-gram "according to" which is the same as the reference.

In the selected data, the only sentence containing the *n*-gram "nach Krankenhausangaben" is the authentic sentence presented in the first row of Table 6 (selected by every execution of TA). As we see, this is a noisy sentence as the target-side does not correspond to an accurate translation (observe that in the source sentence we cannot find names such as "La Passione" or "Carlo Mazzacurati" that are present in the English side). Accordingly, using this sentence in the training of the NMT is harmful.

### 5.2 Analysis of Back-translated Sentences

We find many cases where artificially-generated data is more useful for NMT models than authentic translations. In Table 6 we show some examples.

In rows 1 and 2 we present sentences in which the artificial sentence (*German (synth)* column) is a better translation than the authentic counterpart. In addition to the example described previously (the example of the first row), we also see in row 2 that the authentic candidate pair is ("die Veranstalter haben viele Konzerte und Recitale geplant. Es wird für uns eine vorzügliche Gelegenheit sein Ihre Freizeit angenehm zu gestalten und Sie für die ernste Musik zu gewinnen.","every participant will play at least one programme.") whereas the synthetic counterpart is the pair ("jeder Teilnehmer wird mindestens ein Programm spielen.","every participant will play at least one programme."). In this case, it is preferable to use the synthetic sentence for training instead of the authentic as it is a more accurate translation (observe that the authentic German side consists of two sentences and it is longer than the English-side).

We also present a case in which both authentic and artificial sentences are not proper translations of the English sentence, so both sentences would hurt the performance of NMT if used for training. In row 3 there is a noisy sentence that should not

| | | | batch | | | online | | |
|---|---|---|---|---|---|---|---|---|
| | | FDA | $\gamma = 0.75$ | $\gamma = 0.50$ | $\gamma = 0.25$ | $\gamma = 0.75$ | $\gamma = 0.50$ | $\gamma = 0.25$ |
| BIO | | | | | | | | |
| 100K lines | BLEU | 33.68 | 33.56 | 33.48 | 33.3 | 33.39 | 33.41 | 33.45 |
| | TER | 45.97 | 46.51 | 46.49 | 46.93 | 46.33 | 46.58 | 46.7 |
| | METEOR | 34.71 | **34.97** | **35.09*** | **34.89** | **34.97** | **35.00** | **35.11*** |
| | CHRF3 | 59.24 | **59.48** | **59.62** | **59.59** | **59.45** | **59.55** | **59.83*** |
| 200K lines | BLEU | 33.96 | 33.75 | 33.86 | 32.96 | 33.93 | 33.54 | 33.33 |
| | TER | 45.64 | 46.17 | 46.14 | 47.04 | 45.99 | 46.26 | 46.77 |
| | METEOR | 35.01 | **35.08** | **35.08** | 34.96 | 35.01 | **35.09** | 34.98 |
| | CHRF3 | 59.56 | **59.58** | **59.79** | **59.77** | **59.63** | **59.85** | **59.65** |
| 500K lines | BLEU | 33.75 | 33.72 | 33.74 | 33.13 | 33.75 | **33.95** | 33.46 |
| | TER | 45.92 | 46.32 | 46.13 | 46.90 | 46.12 | 46.14 | 46.85 |
| | METEOR | 34.92 | **35.15** | **35.16** | 34.87 | **35.14** | **35.11** | **34.93** |
| | CHRF3 | 59.57 | 59.84 | 59.83 | 59.61 | **59.83** | **60.00*** | **59.77** |
| NEWS | | | | | | | | |
| 100K lines | BLEU | 26.49 | **26.58** | **26.61** | **26.60** | **26.51** | **26.59** | **26.54** |
| | TER | 54.21 | 54.34 | 54.69 | 54.94 | 54.26 | 54.38 | 54.8 |
| | METEOR | 30.21 | **30.36*** | **30.43*** | **30.51*** | **30.33*** | **30.49*** | **30.49*** |
| | CHRF3 | 51.80 | **52.09*** | **52.37*** | **52.52*** | **52.06*** | **52.41*** | **52.49*** |
| 200K lines | BLEU | 26.55 | **26.62** | **26.66** | **26.65** | **26.77*** | **26.72** | **26.80*** |
| | TER | 54.17 | 54.36 | 54.52 | 54.70 | 54.29 | 54.48 | 54.55 |
| | METEOR | 30.24 | **30.37*** | **30.49*** | **30.55*** | **30.45*** | **30.54*** | **30.61*** |
| | CHRF3 | 51.89 | **52.14*** | **52.49*** | **52.55*** | **52.35*** | **52.56*** | **52.75*** |
| 500K lines | BLEU | 26.40 | **26.70*** | **26.92*** | **26.73** | **26.68*** | **26.94*** | **26.98*** |
| | TER | 54.47 | **54.29** | **54.37** | **54.63** | **54.13*** | 54.35 | 54.56 |
| | METEOR | 30.1 | **30.44*** | **30.59*** | **30.59*** | **30.42*** | **30.58*** | **30.68*** |
| | CHRF3 | 51.71 | **52.29*** | **52.57*** | **52.66*** | **52.19*** | **52.58*** | **52.81*** |

Table 5: Results of the models built with different sizes of FDA-selected hybrid data following the *batch* and *online* approaches. The results in bold indicate an improvement over FDA. The asterisk means the improvement is statistically significant at p=0.01.

have been included as the target side is not English but French. The TAs search for *n*-grams in the source side, so as in this case the artificial sentence consists of a sequence of dots (the BT model has not been able to translate the French sentence) this prevents the TA from selecting it, whereas the authentic sentence-pair could be selected as it is a natural German sentence.

Surprisingly, this correction of inaccurate translations can also be seen on the set of sentences that have been used for training the BT model. As this model does not overfit, when it is provided with the same target sentence used for training, it is capable of generating different valid translations. For example, in row 4 of Table 6 we see the pair ("die Preise liegen zwischen 32.000 und 110.000 Won.","the first evening starts with a big parade of all participants through the city towards the beach.") which is one of the sentence pairs used for training the BT model. This is a noisy sentence (see, for instance, that the English-side does not include the numbers). However, the sentence generated by the BT model is "der erste Abend beginnt mit einer großen Parade aller Teilnehmer durch die Stadt zum Strand." which is a more accurate translation of than the sentence used for training the model that generates it.

## 6 Conclusion and Future Work

In this work, we have presented how artificially generated sentences can be used to augment a set of candidate sentences so data-selection algorithms have a wider variety of sentences to select from. The TA-selected sets have been evaluated

| | German (auth) | German (synth) | English |
|---|---|---|---|
| 1 | nach Krankenhausangaben wurden zwei um die 50 Jahre alte Männer durch das Beben schwer verletzt: einer sei von einem herabfallenden Schornstein getroffen worden, der andere habe durch Glas Schnittwunden erlitten. außerdem seien mehrere Menschen durch herabstürzende Gegenstände in ihren Wohnungen leicht verletzt worden. | am Samstag wird es eine weitere Komödie, "La pasone"von Carlo Mazzacurati Italiens, geben, die die fruchtbare Silvio Orlando, die ein washed-up filmin der Toskana ist, in einer Nachbarkapelle aus dem 16. Jh. | Saturday will feature another comedy, "La Passione"by Carlo Mazzacurati of Italy starring the prolific Silvio Orlando, who plays a washed-up filmmaker who is forced to set his last-chance project in Tuscany after a plumbing disaster at his country home damages a 16th-century fresco in a neighbouring chapel. |
| 2 | die Veranstalter haben viele Konzerte und Recitale geplant. Es wird für uns eine vorzügliche Gelegenheit sein Ihre Freizeit angenehm zu gestalten und Sie für die ernste Musik zu gewinnen. | jeder Teilnehmer wird mindestens ein Programm spielen. | every participant will play at least one programme. |
| 3 | folglich übernimmt Informatic SA keine Gewährleistung für ihre Richtigkeit , ausser sie wurden vom Kunden schriftlich oder per E-Mail ausdrücklich für obligatorisch erklärt . | . . , . . . . . | par conséquent, Informatic SA ne donne donc aucune assurance quant à leur exactitude à moins qu'elles n'aient été expressément déclarées obligatoires par écrit ou par e-mail par le client . |
| 4 | die Preise liegen zwischen 32.000 und 110.000 Won. | der erste Abend beginnt mit einer großen Parade aller Teilnehmer durch die Stadt zum Strand. | the first evening starts with a big parade of all participants through the city towards the beach. |

Table 6: Examples of back-translated sentences

according to how useful they are for improving NMT models.

We have presented three methods of creating such hybrid data: (i) by allowing the TA decide whether to select authentic or synthetic data (*hybr*); (ii) by performing independent executions of the TA on authentic and synthetic sets (*batch*); and (iii) using an MT-generated seed to select monolingual sentences so only the extracted subset is back-translated (*online*).

The experiments showed that artificially-generated sentences can be as competitive as authentic data, as models built with different proportions of authentic and synthetic data achieve similar or even better performance than those fine-tuned with authentic pairs only. On one hand, those sentences whose target-sides could hurt the performance of NMT (such as sentences in a different language to that expected) causes the back-translated sentence to also contain unnatural *n*-grams and so TAs would not select them. On the other hand, if the source-side sentence is not an accurate translation of the target side (the problem of comparable corpus), the back-translated counterpart can be a better alternative to use as training data.

In the future, we want to explore other language pairs and other transductive algorithms. Another limitation of this work is that we have augmented the candidate pool with synthetic sentences generated by a single model. We propose to explore whether using several models for generating the synthetic sentences (including different approaches such as combining statistical and neural model (Poncelas et al., 2019c)) to augment the candidate pool can cause the selected data to further improve NMT models.

## 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).

## References

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72, Ann Arbor, Michigan.

Ergun Biçici. 2013. Feature decay algorithms for fast deployment of accurate statistical machine translation systems. In *Proceedings of the Eighth Work-*

*shop on Statistical Machine Translation*, pages 78–84, Sofia, Bulgaria.

Ergun Biçici and Deniz Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 272–283, Edinburgh, Scotland.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisboa, Portugal.

Mara Chinea-Rios, Alvaro Peris, and Francisco Casacuberta. 2017. Adapting neural machine translation with parallel synthetic data. In *Proceedings of the Second Conference on Machine Translation*, pages 138–147, Copenhagen, Denmark.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. 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 (Volume 2: Short Papers)*, pages 176–181, Portland, Oregon.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Guillem Gascó, Martha-Alicia Rocha, Germán Sanchis-Trilles, Jesús Andrés-Ferrer, and Francisco Casacuberta. 2012. Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 152–161, Avignon, France.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–1780.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 67–72, Vancouver, Canada.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. One sentence one model for neural machine translation. *arXiv preprint arXiv:1609.06490*.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79, Da Nang, Vietnam.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Zuzanna Parcheta, Germán Sanchis-Trilles, and Francisco Casacuberta. 2018. Data selection for nmt using infrequent n-gram recovery. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 219–227, Alacant, Spain.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2018a. Data selection with feature decay algorithms using an approximated target side. In *15th International Workshop on Spoken Language Translation*, pages 173–180, Bruges, Belgium.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2018b. Feature decay algorithms for neural machine translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 239–248, Alacant, Spain.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2019a. Adaptation of machine translation models with back-translated data using transductive data selection methods. In *20th International Conference on Computational Linguistics and Intelligent Text Processing*, La Rochelle, France.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2019b. Transductive data-selection algorithms for fine-tuning neural machine translation. In *Proceedings of The 8th Workshop on Patent and Scientific Literature Translation*, pages 13–23, Dublin, Ireland.

Alberto Poncelas, Maja Popovic, Dimitar Shterionov, Gideon Maillette de Buy Wenniger, and Andy Way. 2019c. Combining SMT and NMT back-translated data for efficient NMT. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 922–931, Varna, Bulgaria.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018c. Investigating backtranslation in neural machine translation. In *21st Annual Conference of the European Association for Machine Translation*, pages 249–258, Alacant, Spain.

Alberto Poncelas, Andy Way, and Kepa Sarasola. 2018d. The ADAPT system description for the IWSLT 2018 Basque to English translation task. In *15th International Workshop on Spoken Language Translation*, pages 76–82, Bruges, Belgium.

Maja Popovic. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1715–1725, Berlin, Germany.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. 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*, pages 223–231, Cambridge, Massachusetts, USA.

Vladimir N. Vapnik. 1998. *Statistical Learning Theory*. Wiley-Interscience.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondrej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247, Copenhagen, Denmark.