EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION

EAMT2018

Proceedings of the

# 21st Annual Conference of the European Association for Machine Translation

28–30 May 2018

Universitat d'Alacant

Alacant, Spain

***Edited by***

Juan Antonio Pérez-Ortiz

Felipe Sánchez-Martínez

Miquel Esplà-Gomis

Maja Popović

Celia Rico

André Martins

Joachim Van den Bogaert

Mikel L. Forcada

***Organised by***

Universitat d'Alacant

Universidad de Alicante

transducens

research group


**ISBN:** 978-84-09-01901-4

# Investigating Backtranslation in Neural Machine Translation

**Alberto Poncelas, Dimitar Shterionov, Andy Way,**
**Gideon Maillette de Buy Wenniger** and **Peyman Passban**
School of Computing, DCU, ADAPT Centre
{firstname.lastname}@adaptcentre.ie

## Abstract

A prerequisite for training corpus-based machine translation (MT) systems – either Statistical MT (SMT) or Neural MT (NMT) – is the availability of high-quality parallel data. This is arguably more important today than ever before, as NMT has been shown in many studies to outperform SMT, but mostly when large parallel corpora are available; in cases where data is limited, SMT can still outperform NMT.

Recently researchers have shown that back-translating monolingual data can be used to create synthetic parallel corpora, which in turn can be used in combination with authentic parallel data to train a high-quality NMT system. Given that large collections of new parallel text become available only quite rarely, backtranslation has become the norm when building state-of-the-art NMT systems, especially in resource-poor scenarios.

However, we assert that there are many unknown factors regarding the actual effects of back-translated data on the translation capabilities of an NMT model. Accordingly, in this work we investigate how using back-translated data as a training corpus – both as a separate standalone dataset as well as combined with human-generated parallel data – affects the performance of an NMT model. We use incrementally larger amounts of back-translated data to train a range of NMT systems for German-to-English, and analyse the resulting translation performance.

## 1 Introduction

Neural Machine Translation (NMT) [Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015] is a relatively new machine translation (MT) paradigm that has quickly become dominant in both academic and industry MT communities, achieving state-of-the-art results [Bentivogli et al., 2016; Bojar et al., 2016; Junczys-Dowmunt et al., 2016; Wu et al., 2016; Castilho et al., 2017; Shterionov et al., 2017] on a range of language pairs and domains. As a corpus-based paradigm, the translation quality strongly depends on the quality and quantity of the training data provided. In comparison to statistical machine translation (SMT) [Koehn, 2010], NMT typically requires more data to build a system with good translation performance [Koehn and Knowles, 2017].

In many use-cases, however, the amount of good-quality parallel data available is insufficient to reach the translation standard required. In such cases, it has become the norm to resort to back-translating freely available monolingual data [Sennrich et al., 2016b; Belinkov and Bisk, 2017; Domhan and Hieber, 2017] to create an additional synthetic parallel corpus [Sennrich et al., 2016b] for training an NMT model.

In this paper, we assert that this scenario has become the default in NMT without proper consideration of the merits of the approach. For example, Rarrick et al. [2011] present an algorithm for filtering noisy content from Web-scraped parallel corpora, in order to mitigate the "pollut[ion] [of the Web] with increasing amounts of machine-translated content". They note that their algorithm "is capable of identifying machine-



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. 249–258
Alacant, Spain, May 2018.

translated content in parallel corpora for a variety of language pairs, and that in some cases it can be very effective in improving the quality of an MT system ... thus challenging the conventional wisdom in natural language processing that 'more data is better data'". Note too that Somers [2005] demonstrates backtranslation (or 'round trip' translation) to be an untrusted means of MT evaluation. In the same vein, Way [2013] notes that in order to show that MT is error-prone, "sites like Translation Party (`http://www.translationparty.com/`) have been set up to demonstrate that continuous use of 'back translation' – that is, start with (say) an English sentence, translate it into (say) French, translate that output back into English, ad nauseum – ends up with a string that differs markedly from that which you started out with".

Surely, then, no-one would argue that building an MT system – whether it be SMT or NMT – with *solely* synthetic data is a good idea; after all, the premise underpinning the paper by Rarrick et al. [2011] was that adding machine-translated data to high-quality human-translated training data *harms* performance. Nonetheless, NMT developers have been seduced into using back-translated data as a means of necessity; there is simply not enough authentic human-translated parallel data available to obtain high-quality results in all scenarios where we would like to deploy NMT. Somewhat surprisingly, despite the inherent problems noted above, adding back-translated data does help improve the quality of NMT output!

In this paper we set out to systematically test from the ground up the merits of back-translated data. We investigate three scenarios: (i) NMT systems trained on 'perfect' human-translated (authentic) data; (ii) using only back-translated (synthetic) data for training NMT systems; and (iii) NMT systems trained on a combination of human-translated and back-translated data. We systematically create multiple training corpora of increasing sizes, using training sets with authentic, synthetic and hybrid (authentic + synthetic) data.

For the hybrid case we increment the back-translated to human-generated data ratio and observe the quality of the resulting NMT systems. We aim to identify to what extent adding synthetic data improves (or harms) the translation capabilities of NMT systems. That is, we investigate whether backtranslation as a core technique in NMT has any limits; given that synthetic data is generated via another imperfect MT system, we hypothesise that NMT trained with 'imperfect' data will – at some point – undo any benefits from the 'perfect' (human-translated) data, and lead the NMT to degrade in performance.[1]

In all our experiments, we exploit data that is widely used in the academic community for researching the quality of MT. The datasets that we use in our experiments all come from the Translation Task of the Tenth Workshop on Machine Translation in 2015 (WMT 2015 [Bojar et al., 2015]).[2] To build our NMT systems we use OpenNMT-py (the pytorch port of OpenNMT [Klein et al., 2017]) with standard settings that allows for easy replicability of our experiments.

The remainder of the paper is structured as follows: Section 2 presents related work on using back-translated and other synthetic data in MT. Section 3 explains how back-translated data affects the training and quality of an NMT system. Our data is described in Section 4, and our experiments are outlined in Section 5. The results are summarised and analysed in Section 6. We conclude in Section 7 with final remarks and future work plans.

## 2 Related Work

Recent studies have shown different approaches to exploiting monolingual data to improve NMT. Gülçehre et al. [2015] present two approaches to integrate a language model trained on monolingual data into the decoder of an NMT system. Similarly, Domhan and Hieber [2017] focus on improving the decoder with monolingual data. While these studies show improved overall translation quality, they require changing the underlying neural network architecture. In contrast, backtranslation allows one to generate a parallel corpus that, consecutively, can be used for training in a standard NMT implementation as presented by Sennrich et al. [2016b]. Sennrich et al. [2016b] use 4.4M sentence pairs of authentic human-translated parallel data to train a baseline English → German NMT system that is later used to translate 3.6M German and 4.2M English target-side sentences. These are then mixed with the initial data to create human + synthetic parallel corpora which are

[1] Note that this should not be confused with the problem of overfitting, where the NMT system learns the training data very well but fails to generalize, with the result that it performs poorly on unseen data.

[2] `http://www.statmt.org/wmt15/`

then used to train new models. Due to the good results that were obtained, adding synthetic data has become a popular step in the NMT training pipeline [Sennrich et al., 2016c; Di Gangi et al., 2017; Lo et al., 2017].

Karakanta et al. [2018] use back-translated data to improve MT for a low-resource language, namely Belarusian (BE). They transliterate a high-resource language (Russian, RU) into their low-resource language (BE) and train a BE→EN system, which is then used to translate monolingual BE data into EN. Finally, an EN→BE system is trained with that back-translated data.

The work of Park et al. [2017] presents an analysis of models trained only with synthetic data. They train NMT models with parallel corpora composed of: (i) synthetic data in the source-side only; (ii) synthetic data in the target-side only; and (iii) a mixture of parallel sentences of which either the source-side or the target-side is synthetic.

Note too that in contrast to the efforts of Rarrick et al. [2011], backtranslation has been applied successfully in PBSMT. Bojar and Tamchyna [2011] use back-translated data to optimize the translation model of a PBSMT system and show improvements in the overall translation quality for 8 language pairs.

## 3 Issues involved in creating back-translated parallel data

Intuitively, MT models built using synthetic data should not perform well. A text translated by a machine can contain errors, so a model trained on such data may learn and replicate these mistakes. While Sennrich et al. [2016b] demonstrated that using back-translated data (in combination with human-translated data) during training can have a positive impact on the performance of the model, we hypothesize that the performance of the model will degrade if the synthetic data is overly dominant in the training set, i.e. the benefit of using high-quality authentic parallel data may be outweighed by the synthetic back-translated data.

We investigate our hypothesis through a systematic analysis of NMT models trained on different-sized parallel datasets containing increasing amounts of back-translated data. We acknowledge the plethora of factors that may impact such an analysis, e.g. vocabulary size, learning optimizer, learning rate, total amount of training steps/minibatches, etc. However, with this work we aim to provide a solid experimental baseline NMT set-up that would facilitate the analysis of the impacts of adding synthetic data to the training corpus. Furthermore, our analysis does not aim to compare the best possible systems, but rather NMT systems trained under the same conditions that would allow a fair comparison. In this regard, we train our systems with word-based dictionaries, rather than with dictionaries based on sub-word units e.g., using Byte-Pair Encoding (BPE) [Sennrich et al., 2016a], although the latter case generally leads to higher MT quality. Given two models of the same size (one trained on authentic and one on synthetic data) the same words can be split into sub-words differently. As such, the quality differences could be due to the sub-word units, learned from the specific data rather than the differences in the authentic and synthetic data.

Our evaluation builds a clearer picture of the progressive effects of adding synthetic data to the training corpus of NMT engines. To the best of our knowledge, such an analysis has not been performed at the time of writing.

Furthermore, we compare NMT systems built on authentic-only data to systems built on synthetic-only data and put the two extremes to a test. We hypothesise that only synthetic data will not be enough to train an NMT system with good performance due to the errors mediated by the initial MT system used to generate that data. However, our results are more than a little surprising. We present detailed analysis of our empirical results in Section 6.

## 4 Data

For the scope of this work, we use the German–English parallel data of the WMT 2015 Translation task [Bojar et al., 2015]. This corpus is shuffled, tokenized, truecased and cleaned (removing sentences of length over 126 words). In total, it contains 4.48M sentence pairs (225M words).

In order to explore the effects of back-translated data, we use human-translated (authentic) and back-translated (synthetic) data in three possible configurations:

- Authentic data only: Models are trained using authentic data only. Such models provide a baseline that any other model can be compared to. This is the baseline scenario for quality of data. Furthermore, such models represent a use-case where an industry partner supplies authen-

tic data to MT engineers in order to build an NMT system.

- Synthetic data Only: Models are built using back-translated data only. Such models represent the case where no parallel data is available but monolingual data can be translated via an existing MT system and provided as a training corpus to a new NMT system. Such cases appear as the other extreme, or the worst-case scenario for quality of data. They reflect resource limitations, either due to the physical unavailability of data, i.e. low-resource languages, or due to economic reasons. Using synthetic data only might also be an option in cases where a high-quality model trained on real data is available, but the translation task is on a very different domain than the training data. In this case using the high-quality model to back-translate domain-specific monolingual target data, and then building a new model with this synthetic training data, might be useful for domain adaptation.
- Hybrid data: Models are built using a base dataset of 1M authentic sentence pairs combined with differing amounts of back-translated data. This is the most interesting scenario (similar to Sennrich et al. [2016b]) which allows us to trace the changes in quality with increases in synthetic-to-authentic data ratio.

All the models that we built are evaluated using the same test set. This test set is provided by WMT 2015 news translation task. It consists of 2169 sentences from the news domain. These sentences have also been tokenized and truecased.

## 5 Experimental set-up

We train sequence-to-sequence NMT models [Sutskever et al., 2014] based on recurrent neural networks with an attention mechanism [Bahdanau et al., 2015; Luong et al., 2015]. The NMT framework we use is OpenNMT [Klein et al., 2017] and in particular its pytorch[3] port.

Our set-up follows the OpenNMT guidelines,[4] that indicate that the default training configuration is reasonable for training a German-to-English model on WMT 2015 data.

We acknowledge the multitude of parameters and values that one can tweak in the set-up of an NMT system, leading to systems with significantly different performance. Moreover, the choice of these parameters often depends on the training data. In our experiments, however, we have focused on a static NMT set-up, where the different parameters (e.g. the NMT learning optimizer, number of epochs, etc.) are common for all systems we train. The decision on our set-up is based on two factors: (i) by limiting the variability of parameters, we can more easily investigate the effects of back-translated data by directly comparing the translation quality of the resulting NMT systems; and (ii) while certain new architectures such as *Transformer* [Vaswani et al., 2017] or different settings might obtain even better results, our goal here is not to build the absolutely best possible systems, but rather use configurations that are representative of what is used in the field and allow easy replication. Specifically, we use a 2-layer LSTM [Hochreiter et al., 1997] with 500 hidden units, a vocabulary size of 50,002 for the source language and 50,004 for the target language. A model is trained for 13 epochs, using the stochastic gradient descent learning optimizer and a batch size of 64. Any unknown words in the translation are replaced with the word in the source language that has the highest attention.

We first trained a baseline $DE \rightarrow EN$ model on $1,000,000$ parallel sentences of authentic data (*base dataset*) and a baseline $EN \rightarrow DE$ model on the same data set with source and target sides swapped around. The latter model is used for back-translation to create *synthetic dataset*s. We found that using 1M sentences to train the model was sufficient for 'good enough' translations. To determine this, we performed preliminary tests that involve human evaluation alongside automatic metrics (on a random sample of the outputs) with models trained on other data sizes.[5] When performing backtranslation, we also replace any unknown words with the word in English (the source language when performing the backtranslation) having the highest attention. We used this engine to then back-translate different portions of our original data set that we then used as parallel training data in two different scenarios: (i) by itself, i.e. synthetic data only, and (ii) in combination with the authentic data used to train the first engine, i.e. the hybrid models, as defined in Section 4.

[3] `http://pytorch.org`

[4] `http://opennmt.net/Models/`

[5] These experiments go beyond the scope of this work and are not included in the current paper.

To make our comparison fair, we defined two cases of authentic data. The first one starts with the first 1,000,000 sentences and grows incrementally (adding 500,000 parallel sentences each time) until it contains 3,500,000 sentences, i.e. ranging between the $1^{st}$ and the 3,500,000$^{th}$ sentence. We denote these sets as *auth*$_{0+}$. The *hybr* data sets are composed of the $1^{st}$ 1,000,000 authentic sentences, combined with back-translated data for each following subset of 500,000 sentences.

In the second case, the authentic data sets start from the 1,000,000$^{th}$ sentence. The first one contains 1,000,000 sentences; the next ones increment with 500,000 additional authentic sentences with the last one ranging between the 1,000,000$^{th}$ to the 4,480,000$^{th}$ sentence. These sets we refer to as *auth*$_{1+}$. The *synth* data sets are simply the back-translated data sets from the *auth*$_{1+}$ category.

In this way we compare engines trained on exactly the same original data – *auth*$_{0+}$ to *hybr* and *auth*$_{1+}$ to *synth* – which in one case has been partially or fully back-translated.

In Table 1 we present the percentage of tokens (words, numbers and other symbols) of the test set that are covered by the vocabularies we use to build our models.

| data size | *auth*$_{0+}$ | *hybr* | *auth*$_{1+}$ | *synthetic* |
|---|---|---|---|---|
| 1M | 67.03% | - | 66.35% | 60.81% |
| 1.5M | 67.15% | 66.14% | 66.44% | 60.93% |
| 2M | 67.11% | 65.10% | 66.41% | 60.97% |
| 2.5M | 67.25% | 64.60% | 66.36% | 61.03% |
| 3M | 67.30% | 64.15% | 66.47% | 60.98% |
| 3.5M | 67.25% | 63.77% | 66.55% | 61.01% |

**Table 1:** Coverage of the vocabularies (the top-50000 words) on the tokens in the test set.

## 6 Results

Tables 2 and 3 show the evaluation scores of the models we trained for the authentic-to-hybrid and authentic-to-synthetic cases, respectively. We use a number of common evaluation metrics – BLEU [Papineni et al., 2002], TER [Snover et al., 2006], METEOR [Banerjee and Lavie, 2005], and CHRF [Popovic, 2015] – to give a more comprehensive estimation of the comparative translation quality. With the exception of TER, the higher the score, the better the translation is estimated to be; for TER, being an error metric, the lower the score, the better the quality. For comparing the models of the same size, we have also computed the statistical significance (marked with an asterisk) using multeval [Clark et al., 2011] for BLEU, TER and METEOR at level p=0.01 using Bootstrap Resampling [Koehn, 2004].

| | | | |
|---|---|---|---|
| 1M lines | | 1M auth. | - |
| | BLEU | 0.2278 | - |
| | TER↓ | 0.5748 | - |
| | METEOR | 0.269 | - |
| | CHRF1 | 48.7336 | - |
| 1.5M lines | | 1.5M auth. | 1M auth. + 0.5M synth. |
| | BLEU↑ | 0.2347 | 0.2378 |
| | TER↓ | 0.5702 | 0.5681 |
| | METEOR↑ | 0.2735 | 0.2751 |
| | CHRF1↑ | 49.2973 | 49.5145 |
| 2M lines | | 2M auth. | 1M auth. +1M synth. |
| | BLEU↑ | 0.2382 | 0.2421 |
| | TER↓ | 0.5646 | 0.5644 |
| | METEOR↑ | 0.2755 | 0.2771 |
| | CHRF1↑ | 49.6164 | 49.6818 |
| 2.5M lines | | 2.5M auth. | 1M auth. + 1.5M synth. |
| | BLEU↑ | 0.2419 | 0.242 |
| | TER↓ | 0.5592 | 0.5622 |
| | METEOR↑ | 0.2786 | 0.2784 |
| | CHRF1↑ | 50.015 | 49.8781 |
| 3M lines | | 3M auth. | 1M auth. + 2M synth. |
| | BLEU↑ | 0.2446 | 0.2442 |
| | TER↓ | 0.5572 | 0.5621 |
| | METEOR↑ | 0.2792 | 0.2785 |
| | CHRF1↑ | 50.1999 | 49.9244 |
| 3.5M lines | | 3.5M auth. | 1M auth. + 2.5M synth. |
| | BLEU↑ | 0.2435 | 0.2413 |
| | TER↓ | 0.5586 | 0.5651 |
| | METEOR↑ | 0.2788 | 0.277 |
| | CHRF1↑ | 50.0785 | 49.584 |

**Table 2:** Results of models using human-translated or authentic data and back-translated or synthetic data from the *auth*$_{0+}$ and *hybr* sets.

In Figures 2 and 1 we illustrate how the BLEU and METEOR scores of our models (trained on authentic, synthetic and hybrid data) change with increases in the training data.

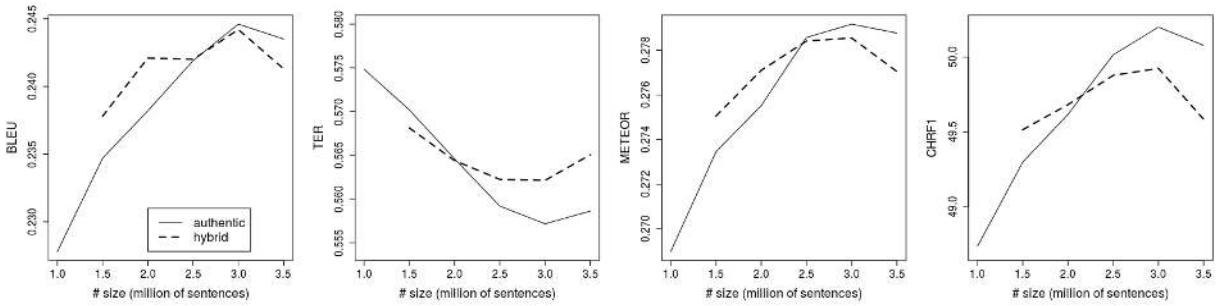


**Figure 1:** Quality scores of NMT systems trained with different sizes of training data from the $auth_{0+}$ and *hybr* sets.

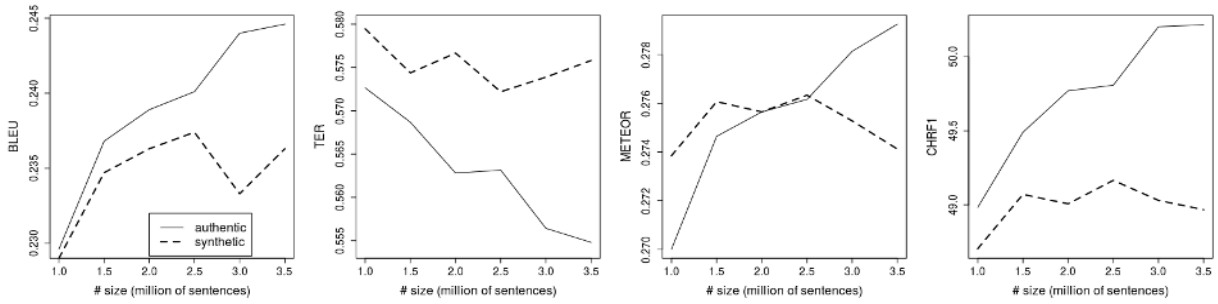


**Figure 2:** Quality scores of NMT systems trained with different sizes of training data from the $auth_{1+}$ and *synth* sets.

### 6.1 Authentic Data Models

In Tables 2 and 3, we see that, as expected, building NMT systems with increasingly larger amounts of human-translated data improves performance: from a BLEU score of 0.2278 with 1M sentence pairs, to the best score of 0.2446 with 3M sentence pairs. This is an absolute improvement of 0.0168, or 7.4% relative. We do, however, see a slight drop when we build our NMT system with 3.5M sentence pairs. All these findings are corroborated by the other three MT evaluation metrics.

### 6.2 Hybrid Data Models

According to the results summarised in Table 2 and Figure 1, the benefits of adding back-translated data presented in Sennrich et al. [2016b] are maintained in our experiments. We see that the hybrid model where 0.5M synthetic sentences are added in the training data (i.e. *1M auth + 0.5M synth* column in Table 2) performs better than the model built with 1M human-translated sentences. In fact, the same-sized hybrid model also outperforms the authentic-only model built with 1.5M sentence pairs.

Adding more and more synthetic data to the training set of an NMT systems causes BLEU scores to rise, as expected, with the best combination comprising 3M sentence pairs (1M authentic and 2M synthetic sentence pairs), which achieves a BLEU score of 0.2442, 0.0066 points absolute better than the smallest hybrid model, a relative improvement of 2.8%.

We see in column *hybr* of Table 1 that the coverage of the hybrid models is not as high as for those built with authentic data only, but in all cases they are higher than for the synthetic-only datasets. We observe that the bigger the data set, the lower the coverage is. We expect that as more synthetic data is added, the more its vocabulary starts to dominate, pushing out words that are more frequent in real parallel data, but less frequent in synthetic data. Accordingly, we expect the coverage of hybrid models to tend to converge to the values of the synthetic models.

Figure 1 shows how the quality of the hybrid models increases the more synthetic data is added. For smaller models, the slopes of the hybrid and authentic models are similar. However, the slope becomes less steep for models trained with 2M sentences or more, as in hybrid datasets with 2M sentence pairs half of it contains synthetic data.

### 6.3 Synthetic Data Models

Earlier in the paper, we suggested that no-one would set out to build an NMT system using solely synthetic data. However, our results show this to

| | | 1M auth. | 1M synth. |
|---|---|---|---|
| 1M lines | BLEU↑ | 0.2296 | 0.2290 |
| | TER↓ | 0.5726* | 0.5795 |
| | METEOR↑ | 0.2700 | 0.2738 |
| | CHRF1↑ | 48.9829 | 48.7035 |
| | | 1.5M auth. | 1.5M synth. |
| 1.5M lines | BLEU↑ | 0.2368* | 0.2347 |
| | TER↓ | 0.5687 | 0.5744 |
| | METEOR↑ | 0.2746 | 0.2761 |
| | CHRF1↑ | 49.4900 | 49.0705 |
| | | 2M auth. | 2M synth. |
| 2M lines | BLEU↑ | 0.2389* | 0.2363 |
| | TER↓ | 0.5628* | 0.5767 |
| | METEOR↑ | 0.2756 | 0.2756 |
| | CHRF1↑ | 49.7702 | 49.0069 |
| | | 2.5M auth. | 2.5M synth. |
| 2.5M lines | BLEU↑ | 0.2401* | 0.2374 |
| | TER↓ | 0.5631* | 0.5722 |
| | METEOR↑ | 0.2762 | 0.2763 |
| | CHRF1↑ | 49.8079 | 49.1656 |
| | | 3M auth. | 3M synth. |
| 3M lines | BLEU↑ | 0.2440* | 0.2333 |
| | TER↓ | 0.5564* | 0.5739 |
| | METEOR↑ | 0.2781* | 0.2753 |
| | CHRF1↑ | 50.2028 | 49.0301 |
| | | 3.5M auth. | 3.5M synth.* |
| 3.5M lines | BLEU↑ | 0.2446* | 0.2363 |
| | TER↓ | 0.5548* | 0.5758 |
| | METEOR↑ | 0.2792* | 0.2741 |
| | CHRF1↑ | 50.2159 | 48.9671 |

**Table 3:** Results of models using human-translated or authentic data and back-translated or synthetic data from the *auth*$_{1+}$ and *synth* sets.

be far from the crazy idea it seemed at the outset (see Table 3 and Figure 2). Using 1M sentence pairs of synthetic-only data (the first of the *synth* data sets), we obtain a BLEU score of 0.229, which continues to rise as we add more synthetic data, achieving the best BLEU score of 0.2363 with 3.5M sentence pairs. This is an absolute improvement of 0.0073, or 3.2% relative. Looking at the other metrics, the picture is rather more mixed; TER, METEOR and CHRF follow a more steady tendency[6].

[6]The only disagreement of BLEU with the rest of the evaluation metrics is the increment in the translation quality of the model trained using 3.5M synthetic sentences (compared to the model trained using 3M synthetic sentences). However this improvement is not statistically significant at level $p = 0.01$.

It is clear, however, that the difference between the quality of engines trained on synthetic and authentic data is rather small. Moreover, the authentic and synthetic data sets of 1,000,000 sentences result in engines where the latter one actually performs better in terms of METEOR. However, even if smaller models built using synthetic data only can perform very close to the level of authentic-only models, it does not appear to be scalable, as the differences in the quality metrics between the two types of engines increase with larger data sizes, i.e. if we look at Figure 2, the quality of the models trained with synthetic data have a relatively lower increase in quality when more back-translated sentences are added.

From column *synth* of Table 1 we notice that the coverage of models built using synthetic data does not increase when more data is added, (all are around 61%). This coverage is much lower than for authentic data models (*auth*$_{1+}$ column), with coverage of more than 66% for all training sizes.

We put this discrepancy in performance down to the limits of the knowledge encoded by the NMT system used for back-translation. In particular, the sentences on the source side are the output of that system, and so (i) the vocabulary of these source-side sentences is always restricted; and (ii) these sentences will contain errors mediated by the initial NMT system. Given enough data, it will reach a steady point and not improve further. We observe this in Figure 2. We can thus conclude that an NMT system trained on synthetic-only data can learn very well the knowledge encoded by the original system used for back-translation, and can even exceed its quality.

It is worth mentioning that models trained on synthetic or on hybrid data outperform the authentic-only models in the lower-sized training data sets. This indicates that in low-resource scenarios it makes sense to exploit back-translation in order to achieve a better NMT system. However, with synthetic-only data, at a given point the performance of the NMT system plateaus, while in the case of hybrid data the quality starts degrading as the synthetic data overpowers the authentic. In our experimental set-up and data we reached this point at a synthetic-to-authentic ratio of 2:1. In the future we will conduct more experiments with different data, data sizes and language pairs, as well as network set-ups to see whether a true tipping point emerges.

We believe this finding will have positive consequences especially for resource-poor scenarios. In particular, we hypothesise that using any existing MT system (or a combination of systems) to translate monolingual data in order to build an NMT system for the intended language direction with that data is likely to result in translation quality similar to that of the initial MT system.

## 7 Conclusion and Future Work

In this work we studied the performance of NMT German-to-English models when incrementally larger amounts of back-translated (or synthetic) data are used for training. We analysed hybrid NMT models built by adding back-translated data to an initial set of human-translated (or authentic) data, and showed that while translation performance tends to improve when larger amounts of synthetic data are added, performance appears to tail off when the balance is tipped too far in favour of the synthetic data; in our experiments we see a drop in performance of 1.2% for the 3.5M hybrid model compared to the 3M hybrid one. We plan to extend these experiments further in our future work, in order to figure out whether there exists a genuine tipping point, i.e. a ratio between the amount of synthetic and authentic data where the model achieves optimal performance, and beyond which the more synthetic data is added, the worse the NMT quality becomes.

We also built models using synthetic data alone. To our surprise, the performance is quite good; the synthetic-only baseline model achieved quality very close to that of the authentic-only engines. Astonishingly, the synthetic-only engine trained with 1M sentences performs better as scored by METEOR than the authentic-only engine trained on the same amount of data.

We believe our findings have important repercussions for resource-poor scenarios, especially where some prior engine – not necessarily an NMT system – exists for the reverse language direction, as this can be used to create arbitrarily large amounts of back-translated data for bootstrapping an NMT engine for the other language direction. We will investigate this further in ongoing work.

In other future work, we also want to explore the effect of adding artificial data to different language pairs and domains. We envisage the current research as the first contribution to an ongoing investigation of the true merits and limits of back-translation. It may well turn out that adding incrementally larger amounts of back-translated data is less harmful than we expect, but at least doing this from the ground up will hopefully result in a set of principles for NMT practitioners, rather than the rather haphazard state of affairs we see before us today.

## 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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*, San Diego, CA, USA, 2015. 15pp.

Satanjeev Banerjee and Alon Lavie. 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, 2005.

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*, 2017.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, USA, 2016.

Ondrej Bojar and Ales Tamchyna. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT@EMNLP 2011*, pages 330–336, Edinburgh, Scotland, 2011.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias

Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation (wmt16). In *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 131–198, Berlin, Germany, 2016.

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. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, 2015.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1):109–120, 2017.

Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D14-1179`.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 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)*, page 176–181, Portland, Oregon, 2011.

Mattia Antonino Di Gangi, Nicola Bertoldi, and Marcello Federico. FBK's participation to the English-to-German News Translation Task of WMT 2017. In *Proceedings of the Second Conference on Machine Translation*, pages 271–275, Copenhagen, Denmark, 2017.

Tobias Domhan and Felix Hieber. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark, 2017.

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

Sepp Hochreiter, Jürgen Schmidhuber, and Corso Elvezia. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. In *Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016. 8pp.

Alina Karakanta, Jon Dehdari, and Josef van Genabith. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32, 2018. 23pp.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, 2017.

Philipp Koehn. 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, 2004.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada, 2017.

Chi-kiu Lo, Boxing Chen, Colin Cherry, George Foster, Samuel Larkin, Darlene Stewart, and Roland Kuhn. NRC Machine Translation System for WMT 2017. In *Proceedings of the Sec-*

*ond Conference on Machine Translation*, pages 330–337, Copenhagen, Denmark, 2017.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, 2015.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

Jaehong Park, Byunggook Na, and Sungroh Yoon. Building a neural machine translation system using only synthetic parallel data. *arXiv preprint arXiv:1704.00253*, 2017.

Maja Popovic. 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, 2015.

Spencer Rarrick, Chris Quirk, and Will Lewis. Mt detection in web-scraped parallel corpora. In *Proceedings of MT Summit XIII*, pages 422–429, Xiamen, China, 2011.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, pages 1715–1725, Berlin, Germany, 2016a.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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, 2016b.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany, 2016c.

Dimitar Shterionov, Pat Nagle, Laura Casanellas, Riccardo Superbo, and Tony O'Dowd. Empirical evaluation of nmt and pbsmt quality for large-scale translation production. In *User track of the 20th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 74–79, Prague, Czech Republic, 2017.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 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, 2006.

Harold Somers. Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop 2005 (ALTW 2005)*, pages 71–77, Sydney, Australia, 2005.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 3104–3112, Montreal, Quebec, Canada, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Thirty-first Annual Conference on Neural Information Processing Systems*, pages 5998–6008, Long Beach, CA., USA, 2017.

Andy Way. Traditional and emerging use-cases for machine translation. In *Proceedings of Translating and the Computer 35*, London, 2013. 12pp.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.