

Combining SMT and NMT Back-Translated Data for Efficient NMT

**Alberto Poncelas, Maja Popović, Dimitar Shterionov,
Gideon Maillette de Buy Wenniger and Andy Way**

School of Computing, DCU, ADAPT Centre

{firstname.lastname}@adaptcentre.ie

Abstract

Neural Machine Translation (NMT) models achieve their best performance when large sets of parallel data are used for training. Consequently, techniques for augmenting the training set have become popular recently. One of these methods is back-translation [21], which consists on generating synthetic sentences by translating a set of monolingual, target-language sentences using a Machine Translation (MT) model.

Generally, NMT models are used for back-translation. In this work, we analyze the performance of models when the training data is extended with synthetic data using different MT approaches. In particular we investigate back-translated data generated not only by NMT but also by Statistical Machine Translation (SMT) models and combinations of both. The results reveal that the models achieve the best performances when the training set is augmented with back-translated data created by merging different MT approaches.

1 Introduction

Machine translation (MT) nowadays is heavily dependent on the quantity and quality of training data. The amount of available good-quality parallel data for the desired domain and/or language pair is often insufficient to reach the required translation performance. In such cases, it has become the norm to resort to back-translating freely available monolingual data as proposed in [21]. That is, one can translate a set of sentences from language L2 into L1 with an already trained MT system for the language pair L2→L1. Then create a synthetic parallel corpus from L1 to L2, with

the source (L1) side being the translated text and the target side being the monolingual data. Back-translation has been shown to be beneficial not only for MT but also for other NLP tasks where data is scarce, e.g. automatic post-editing (APE) [9, 13]. However, the effects of various parameters for creating back-translated (BT) data have not been investigated enough as to indicate what are the optimal conditions in not only creating but also employing such data to train high-quality neural machine translation (NMT) systems.

The work presented in [17] draws an early-stage empirical roadmap to investigating the effects of BT data. In particular, it looks at how the amount of BT data impacts the performance of the final NMT system. In [21] and [17], the systems used to generate the BT data are neural. However, it has been noted that often different paradigms can contribute differently to a given task. For example, it has been shown that applying an APE system based on NMT technology improves statistical machine translation (SMT) output, but has lower impact on NMT output [2, 5].

In this work we assess the impact of different amounts of BT data generated by two different types of MT systems – NMT and SMT. Our contribution is two-fold: (i) we provide a systematic comparison of the BT data by building NMT systems with a combination of SMT and NMT BT data and (ii) we identify the effects of BT data that originates from SMT or NMT on the end-quality of the trained NMT system. We aim to answer the question: "What is the best choice for BT data?"

2 Preparatory Study: the Effect of Back-Translation when Controlling for the Amount of Training Effort

A typical assumption made when training NMT models, is that when more training data is used,

more training effort is warranted. Based on this assumption when training NMT systems what is normally kept constant is the amount of training epochs rather than the amount of training effort in the form of steps/mini-batches. Nevertheless, when adding back-translated data to the training set, while keeping the amount of epochs the same, the effective amount of training increases. It could then be questioned whether the extra training effort in itself does not partly explain the positive effect of back-translation. For this reason, we seek to answer the question: “Does the effect of back-translation change when we control for the amount of training effort, by keeping the total amount of steps/mini-batches constant?”. To answer this question we compare the performance of systems trained on purely authentic data to those trained on authentic plus synthetic data, while keeping either the number of steps/mini-batches or the number of epochs constant in both settings:

1. Models trained with 1M auth + 2M synth sentences using the default settings, including 13 training epochs.
2. Models trained on 1M auth data only, trained either:
 - (a) using the default settings, including 13 training epochs.
 - (b) Trained for 39 epochs, to obtain a same amount of training effort as for the 1M auth + 2M synth sentences model.

When increasing the epochs to 39, we take appropriate measures to keep the starting point and speed of decay of the learning rate constant for the amount of training steps/epochs.¹

The results of these experiments indicate that training a model on authentic data with 1/3 of the amount of the total parallel data (authentic + synthetic) for an additional 26 epochs to account for the extra training effort is not required as no significant improvement has been observed. Based on the outcome of these experiments we chose the rest of our experiments.

¹This is implemented by changing the start of the learning rate decay from epoch 8 to epoch 22 ($= 7 * 3 + 1$) and changing the decay factor from 0.5 to $\sqrt[3]{0.5} = 0.7936$. This way, the learning rate decay starts after the same amount of data when using the 1M auth dataset ($7 \times 3M$) and the decay rate is maintained at 0.5 for each $3M$ sentences from this point onwards.

3 Using Back-Translation from Different Sources

The work of [21] showed that adding BT data is beneficial to achieve better translation performances. In this work we compare the details related to the translation hypotheses originating from SMT and NMT back-translated training data as well as combine the data from those two different sources. To the best of our knowledge, this has not been investigated yet.

We compare German-to-English translation hypotheses generated by systems trained (i) only on authentic data, (ii) only on synthetic data, and (iii) on authentic data enhanced with different types of BT data: SMT, NMT. We exploit two types of synthetic and authentic data combinations: (a) randomly selected half of target sentences back-translated by SMT and another half by NMT system, and (b) joining all BT data (thus repeating each target segment).

The translation hypotheses are compared in terms of four automatic evaluation metrics: BLEU [15], TER [23], METEOR [1] and CHRF [19]. These metrics give an overall estimate of the quality of the translations with respect to the reference (human translation of the test set). In addition, the translation hypotheses are analyzed in terms of five error categories, lexical variety and syntactic variety.

4 Related Work

A comparison between MT models trained with synthetic and with authentic data that originate from the same source has been presented in [17]. They show that while the performances of models trained with both synthetic and authentic data are better than those of models trained with only authentic data, there is a saturation point beyond which the quality does not improve by adding more synthetic data. Nonetheless, models trained only with synthetic (BT) data perform very reasonably, with evaluation scores being close to those of models trained with only authentic parallel data. In fact, when appropriately selected, BT data can be used to enhance NMT models [16].

[7] confirmed that synthetic data can sometimes match the performance of authentic data. In addition, a comprehensive analysis of different methods to generate synthetic source sentences was carried out. This analysis revealed that sampling from the model distribution or noising beam out-

puts out-performs pure beam search, which is typically used in NMT. Their analysis shows that synthetic data based on sampling and noised beam search provides a stronger training signal than synthetic data based on argmax inference.

One of the experiments reported in [4] is comparing performance between models trained with NMT and SMT BT data. The best Moses system [12] is almost as good as the NMT system trained with the same (authentic) data, and much faster to train. Improvements obtained with the Moses system trained with a small training corpus are much smaller; this system even decreases the performance for the out-of-domain test. The authors also investigated some properties of BT data and found out that the back-translated sources are on average shorter than authentic ones, syntactically simpler than authentic ones, and contain smaller number of rare events. Furthermore, automatic word alignments tend to be more monotonic between artificial sources and authentic targets than between authentic sources and authentic targets.

[4] also compared training BT data with authentic data in terms of lexical and syntactic variety, segment length and alignment monotony, however they did not analyze the obtained translation hypotheses. In [24] it is shown that MT systems trained on authentic and on backtranslated data lead to general loss of linguistic richness in their translation hypotheses.

5 Experimental Settings

For the experiments we have built German-to-English NMT models using the Pytorch port of OpenNMT [10]. We use the default parameters: 2-layer LSTM with 500 hidden units. The models are trained for the same number of epochs. As the model trained with all authentic data converges after 13 epochs, we use that many iterations to train the models (we use the same amount of epochs). As optimizer we use stochastic gradient descent (SGD), in combination with learning rate decay, halving the learning rate starting from the 8th epoch.

In order to build the models, all data sets are tokenized and truecased and segmented with Byte-Pair Encoding (BPE) [22] built on the joint vocabulary using 89500 merge operations. For testing the models we use the test set provided in the WMT 2015 News Translation Task [3]. As development set, we use 5K randomly sampled sen-

tences from development sets provided in previous years of WMT.

6 Data

The parallel data used for the experiments has been obtained from WMT 2015 [3]. We build two parallel sets with these sentences: *base* (1M sentences) and *auth* (3M sentences). We use the target side of *auth* to create the following datasets:

- *SMTsynth*: Created by translating the target-side sentences of *auth*. The model used to generate the sentences is an SMT model trained with *base* set in the English to German direction. It has been built using the Moses toolkit with default settings, using GIZA++ for word alignment and tuned using MERT [14]). The language model (of order 8) is built with the KenLM toolkit [8] using the German side of *base*.
- *NMTsynth*: Created by translating the target-side sentences of *auth*. The model used to generate the sentences is an NMT model (with the same configuration as described in Section 5 but in the English to German direction) trained with the *base* set.
- *hybrNMTSMT*: Synthetic parallel corpus combining *NMTsynth* and *SMTsynth* sets. It has been built by maintaining the same target side of *auth*, and as source side we alternate between *NMTsynth* and *SMTsynth* each 500K sentences.
- *fullhybrNMTSMT*: Synthetic parallel corpus combining all segments from *NMTsynth* and *SMTsynth* sets (double size, each original target sentence repeated twice with both an NMT and SMT back-translation-generated translation).

7 Experiments

In our experiments, we build models on different portions of the datasets described in Section 6. First, we train an initial NMT model using the *base* data set. Then, in order to investigate how much the models benefit from using synthetic data generated by different approaches, we build models with increasing sizes of data (from the data sets described in Section 6).

The models explored are built with data that ranges from 1M sentences (built with only authentic data from *base* data set) to 4M sentences (consisting on 1M sentences from *base* and 3M sentences generated artificially with different models). We also include the models built with the *fullhybrNMTSMT* set. As this set contains duplicated target-side sentences, the largest model we build contains 7M sentences in total but only 4M distinct target-side sentences.

8 Results

8.1 Controlling the Amount of Training Effort

Table 1 shows the effect of controlling the amount of training effort when using back-translation. It can be observed that increasing the number of epochs from 13 to 39 when using just the 1M base training set does not increase the performance over using just 13 epochs (i.e. not compensating the relatively smaller training set with more epochs), rather it deteriorates it. From these results we conclude that there is no reason to believe that the positive effects of using back-translation is caused by an effectively larger training effort, rather than by the advantage of the larger training set itself. We therefore also conclude that it is reasonable to keep the number of epochs constant across experiments, rather than fixing the amount of training effort as measured by steps/mini-batches, and we do the former throughout the rest of the paper.

8.2 Addition of Synthetic Data from SMT and NMT Models

Table 2 shows the results of the performance of the different NMT models we have built. The subtables indicate the size of the data used for building the models (from 1M to 4M lines). In each column it is indicated whether *base* has been augmented with the *auth*, *SMTsynth*, *NMTsynth*, *hybrNMTSMT*, or *fullhybrNMTSMT* data set.

The results show that adding synthetic data has a positive impact on the performance of the models as all of them achieve improvements when compared to that built only with authentic data *1M base*. These improvements are statistically significant at $p=0.01$ (computed with multeval [6] using Bootstrap Resampling [11]). However, the increases of quality are different depending on the approach followed to create the BT data.

First, we observe that models in which SMT-generated data is added do not outperform the models built with the same size of authentic data. For example, the models built with 4M sentences (1M authentic and 3M SMT-produced sentences, in cell + *3M SMTsynth*) achieve a performance comparable to the model trained with smaller number of sentences of authentic data (such as + *1M auth* cell, 2M sentences).

Models built by using NMT-created data have a better performance than those built with data generated by SMT. When performing a pairwise comparison between models using an equal amount of either SMT or NMT-created data, we observe that the latter models outperform the former by around one BLEU point. In fact, the performance of models using NMT-translated sentences is closer to those built with authentic data, and some *NMTsynth* models produce better translation qualities. This is the case of +*1M NMTsynth* model (according to all evaluation metrics) or +*3M NMTsynth* (according to BLEU).

Our experiments also include the performance of models augmented with a combination of SMT- and NMT-generated data. We see that adding *hybrNMTSMT* data, with one half of the data originating from SMT and the other half from NMT models, have performances similar to those models built on authentic data only. According to some evaluation metrics, such as METEOR, the performance is better than *auth* models when adding 1M or 2M artificial sentences (although none of these improvements are statistically significant at $p=0.01$). For these amount of sentences, it also outperforms those models in which only SMT or only NMT BT data have been included.

The models extended with synthetic data that perform best are *fullhybrNMTSMT* models. Furthermore, they also outperform authentic models when built with less than 4M distinct target-sentences according to BLEU, METEOR (showing statistically significant improvements at $p=0.01$) and CHRF1. Despite that, when using large sizes of data (i.e. adding 3M synthetic sentences) the models built with SMT-generated artificial data have the lowest performances whereas the performance of the other three tends to be similar.

	1M <i>base.</i> - 13 Epochs	1M <i>base.</i> - 39 Epochs-	1M <i>base</i> + 2M <i>NMTsynth</i>
BLEU↑	23.40	23.22	25.44
TER↓	57.23	58.21	55.62
METEOR↑	28.09	27.75	29.47
CHRF1↑	50.66	50.18	52.5

Table 1: Results for experimental procedure validation: checking that it is reasonable to use constant number of epochs, not constant amount of training effort, in the experiments.

		1M <i>base.</i>	-	-	-	-
1M lines	BLEU↑	23.40	-	-	-	-
	TER↓	57.23	-	-	-	-
	METEOR↑	28.09	-	-	-	-
	CHRF1↑	50.66	-	-	-	-
		+ 1M <i>auth</i>	+ 1M <i>SMT-synth</i>	+1M <i>NMT-synth</i>	+ 1M <i>hybrN-MTSMT</i>	+ 2M <i>fullhy-brNMTSMT</i>
2M lines	BLEU↑	24.87	24.38	25.32	25.21	25.34
	TER↓	55.81	56.05	55.66	55.87	55.79
	METEOR↑	29.16	28.93	29.33	29.29	29.47
	CHRF1↑	52.03	51.89	52.25	52.36	52.47
		+ 2M <i>auth.</i>	+ 2M <i>SMT-synth</i>	+ 2M <i>NMT-synth</i>	+ 2M <i>hybrN-MTSMT</i>	+ 4M <i>fullhy-brNMTSMT</i>
3M lines	BLEU↑	25.69	24.58	25.44	25.62	25.94
	TER↓	54.99	55.7	55.62	55.25	55.11
	METEOR↑	29.7	29.02	29.47	29.73	29.97
	CHRF1↑	52.77	52.09	52.5	52.89	53.11
		+ 3M <i>auth</i>	+ 3M <i>SMT-synth</i>	+ 3M <i>NMT-synth</i>	+3M <i>hybrN-MTSMT</i>	+ 6M <i>fullhy-brNMTSMT</i>
4M lines	BLEU↑	25.97	24.65	26.01	25.83	25.86
	TER↓	54.54	55.58	55.33	55.17	54.95
	METEOR↑	29.91	29.26	29.71	29.74	29.88
	CHRF1↑	53.16	52.24	52.87	52.84	53.11

Table 2: Performance of models built with increasing sizes of authentic set (first column) and different synthetic datasets (last four columns). +1M, +2M and +3M indicate the amount of sentences added to the *base* set (1M authentic sentences).

8.3 Further Analysis

In order to better understand the described systems, we carried out more detailed analysis of all translation outputs. We analyzed five error categories: morphological errors, word order, omission, addition and lexical errors, and we compared lexical and syntactic variety of different outputs in terms of vocabulary size and number of distinct POS n-grams. We also analyzed the sentence lengths in different translation hypotheses, however no differences were observed, neither in the average sentence length nor in the distribution of different lengths.

Automatic Error Analysis

For automatic error analysis results, we used Hjer-son [18], an open-source tool based on Levenshtein distance, precision and recall. The results are presented in Table 3.

It can be seen that morphological errors are slightly improved by any additional data, but it is hard to draw any conclusions. This is not surprising given that our target language, English, is not particularly morphologically rich. Nevertheless, for all three corpus sizes, the numbers are smallest for the full hybrid system, being comparable to the results with adding authentic data.

training	error class rates↓				
	morph	order	omission	addition	mistranslation
1M <i>base</i>	2.8	9.8	12.0	4.8	29.1
1M <i>base</i> + 1M <i>auth</i>	2.7	9.5	11.4	4.9	28.2
1M <i>base</i> + 1M <i>SMTsynth</i>	2.8	10.0	11.6	4.8	28.1
1M <i>base</i> + 1M <i>NMTsynth</i>	2.7	9.8	10.9	5.0	28.1
1M <i>base</i> + 1M <i>hybrNMTSMT</i>	2.7	9.6	11.4	5.2	27.7
1M <i>base</i> + 1M <i>fullhybrNMTSMT</i>	2.6	9.5	11.0	5.2	27.8
1M <i>base</i> + 2M <i>auth</i>	2.6	9.6	11.2	4.8	27.7
1M <i>base</i> + 2M <i>SMTsynth</i>	2.7	10.0	11.9	4.5	28.0
1M <i>base</i> + 2M <i>NMTsynth</i>	2.6	9.7	11.1	5.1	27.9
1M <i>base</i> + 2M <i>hybrNMTSMT</i>	2.6	9.6	11.0	5.2	27.6
1M <i>base</i> + 2M <i>fullhybrNMTSMT</i>	2.6	9.6	10.7	5.3	27.4
1M <i>base</i> + 3M <i>auth</i>	2.7	9.8	11.2	4.6	27.6
1M <i>base</i> + 3M <i>SMTsynth</i>	2.7	9.8	11.9	4.6	27.9
1M <i>base</i> + 3M <i>NMTsynth</i>	2.5	9.6	11.3	5.3	27.4
1M <i>base</i> + 3M <i>hybrNMTSMT</i>	2.6	9.5	11.0	5.1	27.6
1M <i>base</i> + 3M <i>fullhybrNMTSMT</i>	2.5	9.7	10.8	4.8	27.7

Table 3: Results of automatic error classification into five error categories: morphological error (morph), word order error (order), omission, addition and mistranslation.

As for word order, adding SMT data is not particularly beneficial since it either increases (1M and 2M) or does not change (3M) this error type. NMT systems alone do not help much either, except a little bit for the 3M corpus. Hybrid systems yield the best results for this error category for all corpus sizes, reaching or even slightly surpassing the result with authentic data.

Furthermore, all BT data are beneficial for reducing omissions, especially hybrid which can be even better than the authentic data result.

As for additions, no systematic changes can be observed, except an increase for all types of BT data. However, it should be noted that this error category is reported not to be very reliable for comparing different MT outputs (see for example [20]).

The mostly affected error category is mistranslations. All types of additional data are reducing this type of errors, especially the hybrid BT data for 1M and 2M, even surpassing the effect of adding authentic data. As for the 3M corpus, the improvement in this error category is similar to the one by authentic data, but the best option is to use NMT BT data alone.

In total, the clear advantage of using hybrid systems can be noted for mistranslations, omissions and word order which is the most interesting category. This error category is augmented by adding

BT SMT data or not affected by adding BT NMT data, but combining two types of data creates beneficial signals in the source text.

Lexical and Syntactic Variety

Lexical and syntactic variety is estimated for each translation hypothesis as well as for the human reference translation. The motivation for this is the observation that machine-translated data is generally lexically poorer and syntactically simpler than human translations or texts written in the original language [24]. We want to see how different or similar our translation hypotheses are in this sense, and also how they relate to the reference.

Lexical variety is measured by vocabulary size (number of distinct words) in the given text, and syntactic variety by number of distinct POS n -grams where n ranges from 1 to 4. The results are shown in Figure 1.

First of all, it can be seen that none of the translation hypotheses reaches the variety of the reference translation (the black line on the top). The difference is even more notable for the syntax, where the differences between translation hypotheses are smaller and the difference between them and the reference is larger than for vocabulary.

Furthermore, it can be seen that for authentic data (thin gray line on the bottom and thick

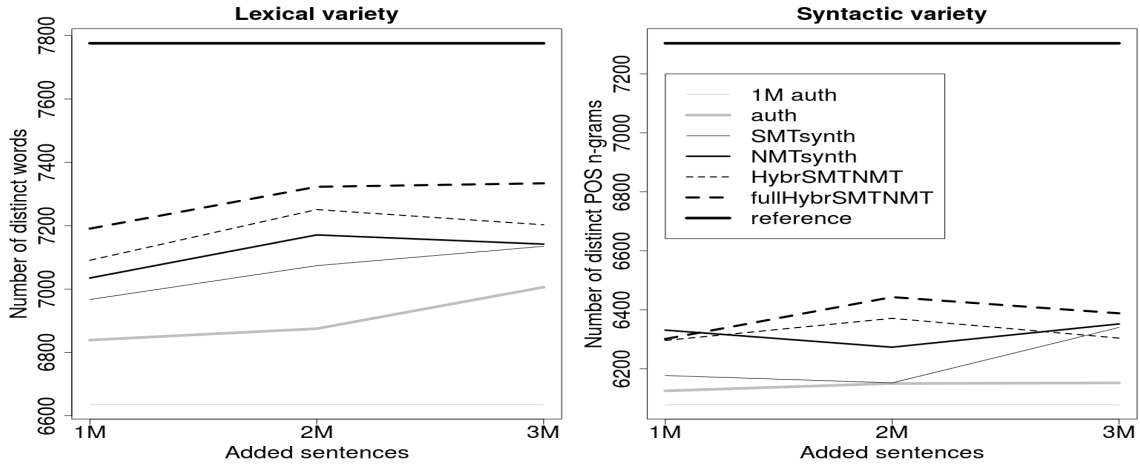


Figure 1: Lexical variety and syntactic variety for all translation hypotheses and for human reference translations.

gray line) the variety increases monotonically with adding more text.

Lexical variety is increased by all synthetic data, too, even more than by authentic data, however, for the NMT and hybrid synthetic data the increase for the 3M corpus is smaller than for smaller corpora.

The increase of syntactic variety is lower both for authentic and for synthetic data than the increase of lexical variety. For 1M and 2M corpus, syntactic variety is barely increased by SMT synthetic data whereas NMT and hybrid data are adding more new instances. For the 3M corpus, however, all synthetic methods yield similar syntactic variety, larger than the one obtained by adding authentic data.

Word/POS 4-gram Precision and Recall

Whereas the increase of lexical and syntactic varieties is a positive trend in general, there is no guarantee that the MT systems are not introducing noise thereby. To estimate how many of added words and POS sequences are sensible, we calculate precision and recall of word and POS 4-grams when compared to the given reference translation. The idea is to estimate how much the translation hypotheses are getting closer to the reference. We take word 4-grams instead of single words because it is not only important that a word makes sense in isolation, but also in a context. Of course, it is still possible that some of the new instances are valid despite being different from the given single reference.

The results of precision and recall for word/POS 4-grams are shown in Figure 2. Several ten-

dencies can be observed:

- hybrid BT data is especially beneficial for the 1M and 2M additional corpora, for 1M even outperforming the authentic additional data, especially regarding word 4-grams;
- NMT BT is the best synthetic option for the 3M additional corpus, however not better than adding 3M of authentic data. This tendency is largest for POS 4-gram precision.
- SMT BT data achieves the lowest scores, especially for POS 4-grams; this is probably related to the fact that it produces less grammatical BT sources, which are then propagated to the translation hypotheses. The differences are largest for the 3M additional corpus, which is probably the reason of diminished effect of the hybrid BT data for this setup.

Overall tendencies are that the hybrid BT data is capable even of outperforming the same amount of authentic data if the amount of added data does not exceed the double size of the baseline authentic data. For larger data, a deterioration can be observed for the SMT BT data, leading to saturation of hybrid models.

Further work dealing with mixing data techniques is necessary, in order to investigate refined selection methods (for example, removing SMT segments which introduce noise).

9 Conclusion and Future Work

In this work we have presented a comparison of the performance of models trained with increasing

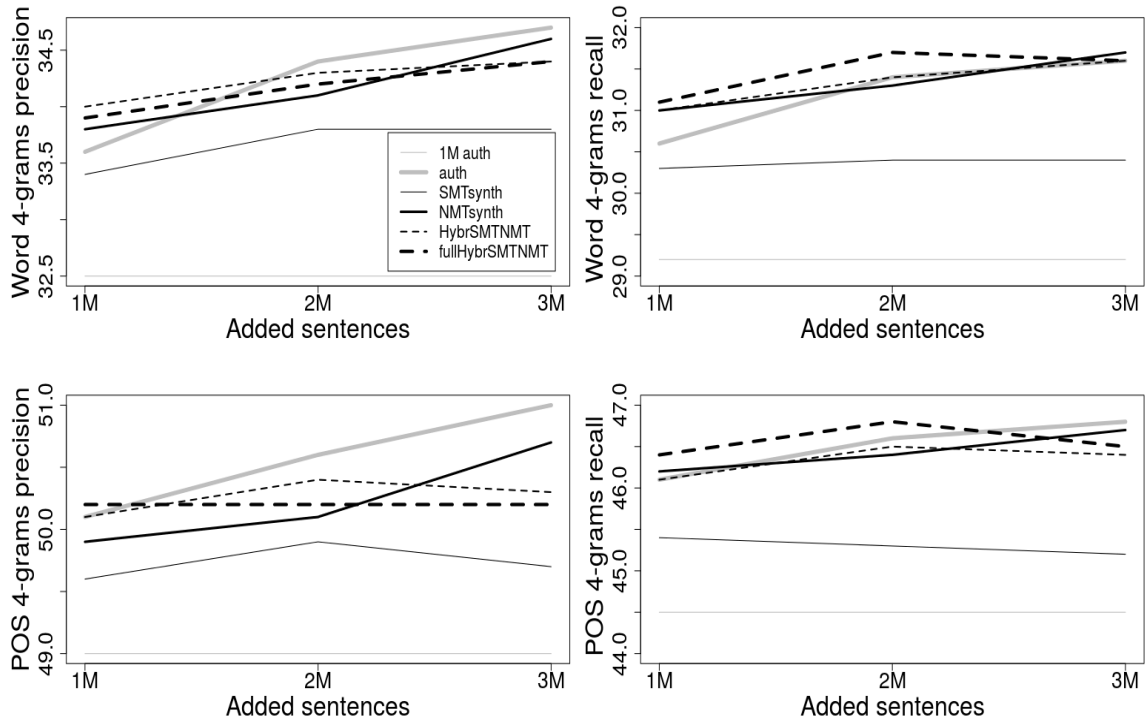


Figure 2: Word/POS 4-gram precision and recall for all translation hypotheses.

size of back-translated data. The artificial data sets explored include sentences generated by using an SMT model, and NMT model and a combination of both. Two mixing strategies are explored: randomly selecting one half of the source segments from the SMT BT data and the other half from the NMT BT data, and using all BT source segments thus repeating each target segment.


Some findings from previous work [4] are confirmed, namely that in terms of overall automatic evaluation scores, SMT BT data reaches slightly worse performance than NMT BT data. Our main findings are that mixing SMT and NMT BT data further improves over each data used alone, especially if full hybridisation is used (using two sources for each target side). These data can even reach better performance than adding the same amount of authentic data, mostly by reducing the number of mistranslations, and increasing the lexical and syntactic variety in a positive way (introducing useful new instances).

However, if the amount of synthetic data becomes too large (three times larger than the authentic baseline data), the benefits of hybrid system start to diminish. The most probable reason is the decrease in grammaticality introduced by SMT BT data which becomes dominant for the larger synthetic corpora.

The presented findings offer several directions for the future work, such as exploring efficient strategies for mixing SMT and NMT data for different authentic/synthetic ratios and investigating morphologically richer target languages.

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] S. Banerjee and A. 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.
- [2] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, et al. Findings of the 2017 conference on machine translation (wmt17). In *Proceed-*

- ings of the *Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark, 2017.
- [3] O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, M. Huck, C. Hokamp, P. Koehn, V. Logacheva, C. Monz, M. Negri, M. Post, C. Scarton, L. Specia, and M. Turchi. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisboa, Portugal, 2015.
- [4] F. Burlot and F. Yvon. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155. Association for Computational Linguistics, 2018.
- [5] R. Chatterjee, M. Negri, R. Rubino, and M. Turchi. Findings of the WMT 2018 shared task on automatic post-editing. In *WMT (shared task)*, pages 710–725. Association for Computational Linguistics, 2018.
- [6] J. H. Clark, C. Dyer, A. Lavie, and N. 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.
- [7] S. Edunov, M. Ott, M. Auli, and D. Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500. Association for Computational Linguistics, 2018.
- [8] K. Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, 2011.
- [9] M. Junczys-Dowmunt and R. Grundkiewicz. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation, WMT 2016, collocated with ACL*, pages 751–758, Berlin, Germany, 2016.
- [10] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 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, 2017.
- [11] P. 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.
- [12] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for SMT. In *Proceedings of 45th annual meeting of the ACL on interactive poster & demonstration sessions*, pages 177–180, Prague, Czech Republic, 2007.
- [13] M. Negri, M. Turchi, R. Chatterjee, and N. Bertoldi. ESCAPE: a large-scale synthetic corpus for automatic post-editing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC*, Miyazaki, Japan, 2018.
- [14] F. Och. Minimum error rate training in statistical machine translation. In *ACL-2003: 41st Annual Meeting of the Association for Computational Linguistics, Proceedings*, pages 160–167, Sapporo, Japan, 2003.
- [15] K. Papineni, S. Roukos, T. Ward, and W.-J. 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, July 2002.
- [16] A. Poncelas, G. M. de Buy Wenniger, and A. Way. 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, 2019.
- [17] A. Poncelas, D. Shterionov, A. Way, G. M. de Buy Wenniger, and P. Passban. Investigating Back translation in Neural Machine Translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*, Alicante, Spain, May 2018.
- [18] M. Popović. Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output. *Prague Bulletin of Mathematical Linguistics*, 96:59–68, October 2011.
- [19] M. Popović. 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.
- [20] M. Popović and A. Burchardt. From human to automatic error classification for machine translation output. In *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT 2011)*, Leuven, Belgium, May 2011.
- [21] R. Sennrich, B. Haddow, and A. 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, 2016.
- [22] R. Sennrich, B. Haddow, and A. 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)*, volume 1, pages 1715–1725, Berlin, Germany, 2016.

- [23] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. 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.
- [24] E. Vanmassenhove, D. Shterionov, and A. Way. Loss and Decay of Linguistic Richness in Neural and Statistical Machine Translation. In *Proceedings of the 17th Machine Translation Summit (MTSummit 2019)*, Dublin, Ireland, August 2019.