Iterative Back-Translation for Neural Machine Translation

Cong Duy Vu Hoang
Computing and Information Systems
University of Melbourne, Australia
vhoang2@student.unimelb.edu.au

Philipp Koehn
Computer Science Department
Johns Hopkins University, USA
phi@jhu.edu

Gholamreza Haffari
Faculty of Information Technology
Monash University, Australia
gholamreza.haffari@monash.edu

Trevor Cohn
Computing and Information Systems
University of Melbourne, Australia
tcohn@unimelb.edu.au

Abstract

We present iterative back-translation, a method for generating increasingly better synthetic parallel data from monolingual data to train neural machine translation systems. Our proposed method is very simple yet effective and highly applicable in practice. We demonstrate improvements in neural machine translation quality in both high and low resourced scenarios, including the best reported BLEU scores for the WMT 2017 German↔English tasks.

1 Introduction

The exploitation of monolingual training data for neural machine translation is an open challenge. One successful method is back-translation (Sennrich et al., 2016b), whereby an NMT system is trained in the reverse translation direction (target-to-source), and is then used to translate target-side monolingual data back into the source language (in the *backward* direction, hence the name back-translation). The resulting sentence pairs constitute a synthetic parallel corpus that can be added to the existing training data to learn a source-to-target model. Figure 1 illustrates this idea.

In this paper, we show that the quality of back-translation matters and propose *iterative back-translation*, where back-translated data is used to build better translation systems in forward and backward directions, which in turn is used to re-back-translate monolingual data. This process can be "iterated" several times. This is a form of co-training (Blum and Mitchell, 1998) where the two models over both translation directions can be used to train one another. We show that iterative back-translation leads to improved results over simple back-translation, under both high and low resource conditions, improving over the state of the art.

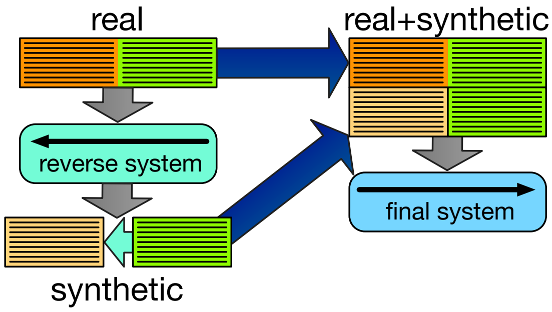

Figure 1: Creating a synthetic parallel corpus through back-translation. First, a system in the reverse direction is trained and then used to translate monolingual data from the target side backward into the source side, to be used in the final system.

2 Related Work

The idea of back-translation dates back at least to statistical machine translation, where it has been used for semi-supervised learning (Bojar and Tamchyna, 2011), or self-training (Goutte et al., 2009, ch.12, p.237). In modern NMT research, Sennrich et al. (2017) reported significant gains on the WMT and IWSLT shared tasks. They showed that even simply duplicating the monolingual target data into the source was sufficient to realise some benefits. Currey et al. (2017) reported similar findings for low resource conditions, showing that even poor translations can be beneficial. Gwinnup et al. (2017) mention in their system description iteratively applying back-translation, but did not report successful experiments.

A more refined idea of back-translation is the *dual learning* approach of He et al. (2016) which integrates training on parallel data and training on monolingual data via round-tripping. We have to admit that we extensively experimented with

Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pages 18–24
Melbourne, Australia, July 20, 2018. ©2018 Association for Computational Linguistics

	English-German	English-French$_{100K}$	English-French$_{1M}$	English-Farsi
parallel en	141 280 704	2 651 040	26 464 159	2 233 688
parallel l_2	134 638 256	2 962 318	29 622 370	2 473 608
mono en	322 529 936	2 154 175 053	2 154 175 053	2 154 175 053
mono l_2	301 736 163	766 646 932	766 646 932	65 585 281

Table 1: Parallel and monolingual corpora used, including English-German, English-French and English-Farsi. Numbers denote the number of words, and l_2 is the second language in each pair. The de-en data is from WMT 2017 (parallel) and a subset of News 2016 (monolingual).

an implementation of this approach, but did not achieve any gains.

An alternative way to make use of monolingual data is the integration of a separately trained language model into the neural machine translation architecture (Gülçehre et al., 2015), but this has not yet to be proven to be as successful as back-translation.

Lampple et al. (2018) explore the use of back-translated data generated by neural and statistical machine translation systems, aided by denoising with a language model trained on the target side.

3 Impact of Back-Translation Quality

Our work is inspired by the intuition that a better back-translation system will lead to a better synthetic corpus, hence producing a better final system. To empirically validate this hypothesis and measure the correlation between back-translation system quality and final system quality, we use a set of machine translation systems of differing quality (trained in the reverse "back-translation" direction), and check how this effects the final system quality.

We carried out experiments on the high-resource WMT German↔English news translation tasks (Bojar et al., 2017). For these tasks, large parallel corpora are available from related domains.[1] In addition, in-domain monolingual news corpora are provided as well, in much larger quantities. We sub-sampled the 2016 news corpus (see Table 1) for about twice as large as corpus as the parallel training corpus.

Following Sennrich et al. (2016b), a synthetic parallel corpus is created from the in-domain news monolingual data, in equal amounts to the existing real parallel corpus. The systems used to translate the monolingual data are canonical attentional neural machine translation systems (Bahdanau et al., 2015). Our setup is very similar to Edinburgh's submission to the WMT 2016 evaluation campaign (Sennrich et al., 2016a),[2] but uses the fast Marian toolkit (Junczys-Dowmunt et al., 2018) for training. We trained 3 different back-translation systems, namely:

10k iterations Training a neural translation model on the parallel corpus, but stopping after 0.15 epochs;

100k iterations As above, but stopping after 1½ epochs; and

convergence As above, but training until convergence (10 epochs, 3 GPU days).

Given these three different systems, we create three synthetic parallel corpora of different quality and train systems on each. Table 2 shows the quality of the final systems. For both direc-

German–English	Back	Final
no back-translation	-	29.6
10k iterations	10.6	29.6 (+0.0)
100k iterations	21.0	31.1 (+1.5)
convergence	23.7	32.5 (+2.9)

English–German	Back	Final
no back-translation	-	23.7
10k iterations	14.5	23.7 (+0.0)
100k iterations	26.2	25.2 (+1.5)
convergence	29.1	25.9 (+2.2)

Table 2: WMT News Translation Task English↔German, reporting cased BLEU on newstest2017, evaluating the impact of the quality of the back-translation system on the final system. Note that the back-translation systems run in the opposite direction and are not comparable to the numbers in the same row.

[1] EU Parliament Proceedings, official EU announcements, news commentaries, and web crawled data.

[2] With true-casing and 50,000 BPE operations (Sennrich et al., 2016c) as pre-processing steps.

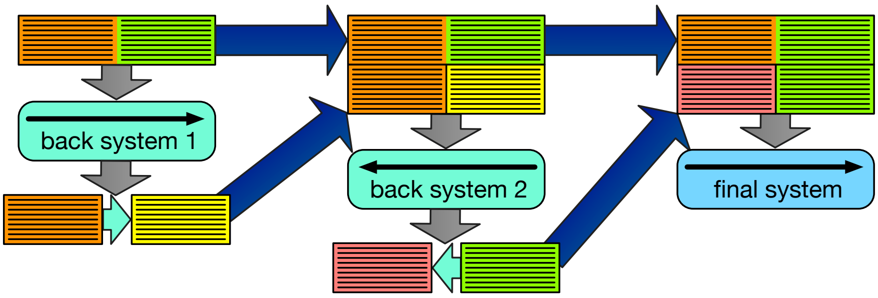

Figure 2: Re-Back-Translation: Taking the idea of back-translation one step further. After training a system with back-translated data (back system 2 above), it is used to create a synthetic parallel corpus for the final system.

tions, the quality of the back-translation systems differs vastly. The **10k iteration** systems perform poorly, and their synthetic parallel corpus provides no benefit over a baseline that does not use any back-translated data.

The longer trained systems have much better translation quality, and their synthetic parallel corpora prove to be beneficial. The back-translation system trained for 100k iteration already provides tangible benefits (+1.5 BLEU for both directions), while the converged system yields even bigger improvements (+2.9 for German–English, and +2.2 for English–German). These results indicate that the quality of the back-translation system is a significant factor for the success of the approach.

4 Iterative Back-Translation

We now take the idea of back-translation one step further. If we can build a better system with the back-translated data, then we can continue repeating this process: Use this better system to back-translate the data, and use this data in order to build an even better system. See Figure 2 for an illustration of this *re-back-translation* process (repeated back-translation). See Algorithm 1 for the details of this *iterated* back-translation process. The final system benefits from monolingual data in both the source and target languages.

We do not have to stop at one iteration of repeated back-translation. We can iterate training the two back-translation systems multiple times. We refer this process to *iterative back-translation.*

In our experiments, we validate our approach under both high-resource and low-resource conditions. Under high-resource conditions, we improve the state of the art with re-back-translation. Under low-resource conditions, we demonstrate the effectiveness of iterative back-translation.

Algorithm 1 Iterative Back-Translation

```
Input: parallel data $D^p$, monolingual source, $D^s$,
    and target $D^t$ text
Let $T_{\leftarrow} = D^p$
repeat
    Train target-to-source model $\Theta_{\leftarrow}$ on $T_{\leftarrow}$
    Use $\Theta_{\leftarrow}$ to create $S = \{(\hat{s}, t)\}$, for $t \in D^t$
    Let $T_{\rightarrow} = D^p \cup S$
    Train source-to-target model $\Theta_{\rightarrow}$ on $T_{\rightarrow}$
    Use $\Theta_{\rightarrow}$ to create $S^{'} = \{(s, \hat{t})\}$, for $s \in D^s$
    Let $T_{\leftarrow} = D^p \cup S'$
until convergence condition reached
Output: newly-updated models $\Theta_{\leftarrow}$ and $\Theta_{\rightarrow}$
```

4.1 Experiments on High Resource Scenario

In §3 we demonstrated that the quality of the back-translation system has significant impact on the effectiveness of the back-translation approach under high-resource data conditions such as WMT 2017 German–English. Here we ask: how much additional benefit can be realised for repeating this process? Also, do the gains for state-of-the-art systems that use deeper models, i.e., more layers in encoder and decoder (Miceli Barone et al., 2017) still apply in this setting?

We evaluate on German–English and English–German, under the same data conditions as in Section 3. We experiment with both *shallow* and *deep* stacked-layer encoder/decoder architectures.

The base translation system is trained on the parallel data only. We train a shallow system using 4-checkpoint ensembling (Chen et al., 2017). The system is used to translate the monolingual data using a beam size of 2.

The first back-translation system is trained on

German–English	**Back***	**Shallow**	**Deep**	**Ensemble**
back-translation	23.7	32.5	35.0	35.6
re-back-translation	27.9	33.6	36.1	36.5
Best WMT 2017	-	-	-	35.1

English–German	**Back***	**Shallow**	**Deep**	**Ensemble**
back-translation	29.1	25.9	28.3	28.8
re-back-translation	34.8	27.0	29.0	29.3
Best WMT 2017	-	-	-	28.3

Table 3: WMT News Translation Task German–English, comparing the quality of different back-translation systems with different final system architectures. *Note that the quality for the back-translation system (Back) is measured in the opposite language direction.

the parallel data and the synthetic data generated by the base translation system. For better performance, we train a deep model with 8-checkpoint ensembling; again we use a beam size of 2.

The final back-translation systems were trained using several different systems: a shallow architecture, a deep architecture, and an ensemble system of 4 independent training runs.

Across the board, the final systems with re-back-translation outperform the final systems with simple back-translation, by a margin of 0.5–1.1 BLEU.

Notably, the final deep systems trained by re-back-translation outperform the state-of-the-art established at the WMT 2017 evaluation campaign for these language pairs, by a margin of about 1 BLEU point. These are the best published results for this dataset, to the best of our knowledge.

Experimental settings For the experiments in the German–English high-resource scenario, we used the Marian toolkit (Junczys-Dowmunt et al., 2018) for training and for back-translation. The shallow systems (also used for the back-translation step) match the setup of Edinburgh's WMT 2016 system (Sennrich et al., 2016a). It is an attentional RNN (default Marian settings) with drop-out of 0.2 for the RNN parameters, and 0.1 otherwise. Training is smoothed with moving average. It takes about 2–4 days.

The deep system uses matches the setup of Edinburgh's WMT 2017 system (Sennrich et al., 2017). It uses 4 encoder and 4 decoder layers (Marian setting `best-deep`) with LSTM cells. Drop-out settings are the same as above. Decoding during test time is done with a beam size of 12, while back-translation uses only a beam size of 2. This difference is reflected in the reported BLEU score for the deep system after back-translation (35.0 for German–English, 28.3 for English–German) and the score reported for the quality of the back-translation system (34.8 (–0.2) and 27.9 (–0.4), respectively) in Table 3.

For all experiments, the true-casing model and the list of BPE operations is left constant. Both were learned from the original parallel training corpus.

4.2 Experiments on Low Resource Scenario

NMT is a data-hungry approach, requiring a large amount of parallel data to reach reasonable performance (Koehn and Knowles, 2017). In a low-resource setting, only small amount of parallel data exist. Previous work has attempted to incorporate prior or external knowledge to compensate for the lack of parallel data, e.g. injecting inductive bias via linguistic constraints (Cohn et al., 2016) or linguistic factors (Hoang et al., 2016). However, it is much cheaper and easier to obtain monolingual data in either the source or target language. An interesting question is whether the (iterative) back-translation can compensate for the lack of parallel data in such low-resource settings.

To explore this question, we conducted experiments on two datasets: A simulated low-resource setting with English–French, and a more realistic setting with English–Farsi. For the English–French dataset, we used the original WMT dataset, sub-sampled to create smaller sets of 100K and 1M parallel sentence pairs. For English–Farsi, we used the available datasets from LDC and TED Talks, totaling about 100K sentence pairs. For detailed statistics see Table 1.

Following the same experimental setup as in high-resource setting,[3] we obtain similar patterns of improvement of translation quality (Table 4).

Back-Translation Generally, it is our expectation that the back-translation approach still improves the translation accuracy in all language pairs with a low-resource setting. In the English–French experiments, large improvements over the baseline are observed in both directions, with +3.5

[3] The difference here is on the NMT toolkit used — we opted to use Amazon's Sockeye (Hieber et al., 2017). We used Sockeye's default configuration with dropout 0.5.

Setting	French–English		English–French		Farsi–English	English-Farsi
	100K	1M	100K	1M	100K	100K
NMT baseline	16.7	24.7	18.0	25.6	21.7	16.4
back-translation	22.1	27.8	21.5	27.0	22.1	16.7
back-translation iterative+1	22.5	-	22.7	-	22.7	17.1
back-translation iterative+2	22.6	-	22.6	-	22.6	17.2
back-translation (w/ Moses)	23.7	27.9	23.5	27.3	21.8	16.8

Table 4: Low Resource setting: Impact of the quality of the back-translation systems on the benefit of the synthetic parallel for the final system in a low-resource setting. Note that, we reported the single NMT systems in all numbers.

BLEU for English to French and +5.4 for French to English in 100K setting. In 1M setting, we also obtained a similar pattern of BLEU gains, albeit of a smaller magnitude, i.e., +1.4 BLEU for English to French and +3.1 for French to English.[4] Note that the large gains here may be due to the fact that the monolingual data is a similar domain to the test data. Inspections of the resulting translations show that the lexical choice has been improved significantly. In English-Farsi experiments shown in Table 4, we also observed BLEU gains, albeit more modest in size: +0.3 BLEU for English to Farsi and +0.4 for Farsi to English. The smaller gains may be because Farsi translation is much more difficult than French; or a result of the diverse mix of domains in the parallel training data (news with LDC and technical talks with TED) where the domain in monolingual data is entirely news, leading to much lower quality than the other datasets. Measuring the impact of iteratively back-translated data in relation to varying domain mismatch between parallel and monolingual data, is a very interesting problem which we will explore in future work; but is out of the scope for this paper.

Balance of real and synthetic parallel data In all our experiments with back-translation, in order to create synthetic parallel data, a small amount of monolingual data is randomly sampled from the big monolingual data (Table 1). As pointed out by (Sennrich et al., 2016b), the balance between the real and synthetic parallel data matters. However, there is no obvious evidence about the affect of the sample size, hence we further studied this by choosing a ratio between the real and synthetic parallel data. We opt to use different ratio (e.g., 1(real):2(synthetic) and 1(real):3(synthetic)) in our experiments. Our results in Table 5 show that more synthetic parallel data seems to be useful (though not obvious), e.g., gains from 16.7 to 16.9 in English to Farsi and gain from 22.1 to 22.4 in Farsi to English.

English–Farsi	**100K**
back-translation 1:1	16.7
back-translation 1:2	16.8
back-translation 1:3	16.9

Farsi-English	**100K**
back-translation 1:1	22.1
back-translation 1:2	22.4
back-translation 1:3	22.4

Table 5: Weighting amounts of real parallel data (1) with varying amounts of synthetic data (1-3). Larger amounts of synthetic data help.

Iterative back-translation For iterative back-translation, we obtained consistent results with the earlier findings from §4.1. In English–French tasks, we see more than +1 BLEU from a further iteration of back-translations, with little difference between 1 or 2 additional iterations. However, in English–Farsi tasks, gains are much smaller.

Comparison to back-translation with Moses We now consider the utility of creating synthetic parallel data from different sources, e.g., from a phrase-based SMT models produced by Moses (Koehn et al., 2007), a considerably faster and more scalable system than modern NMT techniques. As can be seen in Table 4, this has mixed results, being better for English–French, and worse in English–Farsi, than using neural models, although in all cases the results are not far apart.

[4] All the scores are statistically significant with $p < 0.01$.

Quality of the sampled monolingual data Back-translation is dependent much on the quality of back-translated synthetic data. In our paper, repeating the back translation process in 2-3 times can lead to improved translation. However, this can be different in other language pairs and domains. Also, in our work, we sampled the monolingual data uniformly at random, so sentences may be used more than once in subsequent rounds. Its quite likely that other techniques for data sampling and selection, e.g., non-uniform sampling like transductive selection or active learning - which potentially diversifies the quality and quantity of monolingual data - would lead further improvements in translation performance. We leave this for our future work.

Efficacy on iterative back-translation The efficiency of the NMT toolkits we used (sockeye, marian-nmt) is excellent. Both support batch decoding for fast translation, e.g., with a batch-size of 200 (beam-size 5) marian-nmt can achieve over 5000 words per second on one GPU (less than 1 day for translating 4M sentences)[5]; and also this scales linearly to the number of GPUs we have. Alternatively, we can split the monolingual data into smaller parts and distribute these parts over different GPUs. This can greatly speed up the back-translation process. This leaves the problem of training the model in each iteration, which we do 2-3 times. Overall the computational complexity is not a big deal (even with larger dataset), and the iterative back translation is quite feasible with existing modern GPU servers.

5 Conclusion

We presented a simple but effective extension of the back-translation approach to training neural machine translation systems. We empirically showed that the quality of the back-translation system matters for synthetic corpus creation, and that neural machine translation performance can be improved by iterative back-translation in both high-resource and low-resource scenarios. We show empirically that this works well for both high and low resource conditions. The method is simple but highly applicable in practice.

An important avenue for future work is to unify the various approaches to learning, including back-translation (Sennrich et al., 2016b), iterative back-translation (this work), co-training, and dual learning (He et al., 2016) in a framework which can be trained in an end-to-end manner.

Acknowledgements

Most of this work took part during the 2017 JSALT Summer Workshop and we would like to thank the organizers, including our team leader George Foster for their support.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*.

Ondej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Hugh Chen, Scott Lundberg, and Su-In Lee. 2017. Checkpoint ensembles: Ensemble methods from a single training process. *arXiv preprint arXiv:1710.03282*.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Paper*.

Cyril Goutte, Nicola Cancedda, Marc Dymetman, and George Foster. 2009. *Learning Machine Translation*. The MIT Press.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares,

[5] https://marian-nmt.github.io/features/

Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

Jeremy Gwinnup, Timothy Anderson, Grant Erdmann, Katherine Young, Michaeel Kazi, Elizabeth Salesky, Brian Thompson, and Jonathan Taylor. 2017. The afrl-mitll wmt17 systems: Old, new, borrowed, bleu. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 820–828.

F. Hieber, T. Domhan, M. Denkowski, D. Vilar, A. Sokolov, A. Clifton, and M. Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Cong Duy Vu Hoang, Reza Haffari, and Trevor Cohn. 2016. Improving neural translation models with linguistic factors. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 7–14.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. *arXiv preprint arXiv:1804.00344*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based and neural unsupervised machine translation.

Antonio Valerio Miceli Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep architectures for neural machine translation. In *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Paper*.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The university of edinburgh's neural mt systems for wmt17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. 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)*.