Improving Neural Machine Translation Models with Monolingual Data

Document Version:
Peer reviewed version

Published In:
Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics

Improving Neural Machine Translation Models with Monolingual Data

Rico Sennrich and **Barry Haddow** and **Alexandra Birch**
School of Informatics, University of Edinburgh
{rico.sennrich,a.birch}@ed.ac.uk, bhaddow@inf.ed.ac.uk

Abstract

Neural Machine Translation (NMT) has obtained state-of-the art performance for several language pairs, while only using parallel data for training. Target-side monolingual data plays an important role in boosting fluency for phrase-based statistical machine translation, and we investigate the use of monolingual data for NMT. In contrast to previous work, which combines NMT models with separately trained language models, we note that encoder-decoder NMT architectures already have the capacity to learn the same information as a language model, and we explore strategies to train with monolingual data without changing the neural network architecture. By pairing monolingual training data with an automatic back-translation, we can treat it as additional parallel training data, and we obtain substantial improvements on the WMT 15 task English↔German (+2.8–3.7 BLEU), and for the low-resourced IWSLT 14 task Turkish→English (+2.1–3.4 BLEU), obtaining new state-of-the-art results. We also show that fine-tuning on in-domain monolingual and parallel data gives substantial improvements for the IWSLT 15 task English→German.

1 Introduction

Neural Machine Translation (NMT) has obtained state-of-the art performance for several language pairs, while only using parallel data for training. Target-side monolingual data plays an important role in boosting fluency for phrase-based statisti-

cal machine translation, and we investigate the use of monolingual data for NMT.

Language models trained on monolingual data have played a central role in statistical machine translation since the first IBM models (Brown et al., 1990). There are two major reasons for their importance. Firstly, word-based and phrase-based translation models make strong independence assumptions, with the probability of translation units estimated independently from context, and language models, by making different independence assumptions, can model how well these translation units fit together. Secondly, the amount of available monolingual data in the target language typically far exceeds the amount of parallel data, and models typically improve when trained on more data, or data more similar to the translation task.

In (attentional) encoder-decoder architectures for neural machine translation (Sutskever et al., 2014; Bahdanau et al., 2015), the decoder is essentially an RNN language model that is also conditioned on source context, so the first rationale, adding a language model to compensate for the independence assumptions of the translation model, does not apply. However, the data argument is still valid in NMT, and we expect monolingual data to be especially helpful if parallel data is sparse, or a poor fit for the translation task, for instance because of a domain mismatch.

In contrast to previous work, which integrates a separately trained RNN language model into the NMT model (Gülçehre et al., 2015), we explore strategies to include monolingual training data in the training process without changing the neural network architecture. This makes our approach applicable to different NMT architectures.

The main contributions of this paper are as follows:

- we show that we can improve the machine translation quality of NMT systems by mixing monolingual target sentences into the

training set.

- we investigate two different methods to fill the source side of monolingual training instances: using a dummy source sentence, and using a source sentence obtained via back-translation, which we call synthetic. We find that the latter is more effective.

- we successfully adapt NMT models to a new domain by fine-tuning with either monolingual or parallel in-domain data.

2 Neural Machine Translation

We follow the neural machine translation architecture by Bahdanau et al. (2015), which we will briefly summarize here. However, we note that our approach is not specific to this architecture.

The neural machine translation system is implemented as an encoder-decoder network with recurrent neural networks.

The encoder is a bidirectional neural network with gated recurrent units (Cho et al., 2014) that reads an input sequence $x = (x_1, ..., x_m)$ and calculates a forward sequence of hidden states $(\overrightarrow{h}_1, ..., \overrightarrow{h}_m)$, and a backward sequence $(\overleftarrow{h}_1, ..., \overleftarrow{h}_m)$. The hidden states $\overrightarrow{h}_j$ and $\overleftarrow{h}_j$ are concatenated to obtain the annotation vector h_j.

The decoder is a recurrent neural network that predicts a target sequence $y = (y_1, ..., y_n)$. Each word y_i is predicted based on a recurrent hidden state s_i, the previously predicted word y_{i-1}, and a context vector c_i. c_i is computed as a weighted sum of the annotations h_j. The weight of each annotation h_j is computed through an *alignment model* α_{ij}, which models the probability that y_i is aligned to x_j. The alignment model is a single-layer feedforward neural network that is learned jointly with the rest of the network through back-propagation.

A detailed description can be found in (Bahdanau et al., 2015). Training is performed on a parallel corpus with stochastic gradient descent. For translation, a beam search with small beam size is employed.

3 NMT Training with Monolingual Training Data

In machine translation, more monolingual data (or monolingual data more similar to the test set)

serves to improve the estimate of the prior probability $p(T)$ of the target sentence T, before taking the source sentence S into account. In contrast to (Gülçehre et al., 2015), who train separate language models on monolingual training data and incorporate them into the neural network through shallow or deep fusion, we propose techniques to train the main NMT model with monolingual data, exploiting the fact that encoder-decoder neural networks already condition the probability distribution of the next target word on the previous target words. We describe two strategies to do this: providing monolingual training examples with an empty (or dummy) source sentence, or providing monolingual training data with a synthetic source sentence that is obtained from automatically translating the target sentence into the source language, which we will refer to as *back-translation*.

3.1 Dummy Source Sentences

The first technique we employ is to treat monolingual training examples as parallel examples with empty source side, essentially adding training examples whose context vector c_i is uninformative, and for which the network has to fully rely on the previous target words for its prediction. This could be conceived as a form of dropout (Hinton et al., 2012), with the difference that the training instances that have the context vector dropped out constitute novel training data. We can also conceive of this setup as multi-task learning, with the two tasks being translation when the source is known, and language modelling when it is unknown.

During training, we use both parallel and monolingual training examples in the ratio 1-to-1, and randomly shuffle them. We define an epoch as one iteration through the parallel data set, and resample from the monolingual data set for every epoch. We pair monolingual sentences with a single-word dummy source side *<null>* to allow processing of both parallel and monolingual training examples with the same network graph.[1] For monolingual minibatches[2], we freeze the network parameters of the encoder and the attention model.

One problem with this integration of monolin-

---

[1] One could force the context vector c_i to be 0 for monolingual training instances, but we found that this does not solve the main problem with this approach, discussed below.

[2] For efficiency, Bahdanau et al. (2015) sort sets of 20 minibatches according to length. This also groups monolingual training instances together.

gual data is that we cannot arbitrarily increase the ratio of monolingual training instances, or fine-tune a model with only monolingual training data, because different output layer parameters are optimal for the two tasks, and the network 'unlearns' its conditioning on the source context if the ratio of monolingual training instances is too high.

3.2 Synthetic Source Sentences

To ensure that the output layer remains sensitive to the source context, and that good parameters are not unlearned from monolingual data, we propose to pair monolingual training instances with a synthetic source sentence from which a context vector can be approximated. We obtain these through back-translation, i.e. an automatic translation of the monolingual target text into the source language.

During training, we mix synthetic parallel text into the original (human-translated) parallel text and do not distinguish between the two: no network parameters are frozen. Importantly, only the source side of these additional training examples is synthetic, and the target side comes from the monolingual corpus.

4 Evaluation

We evaluate NMT training on parallel text, and with additional monolingual data, on English↔German and Turkish→English, using training and test data from WMT 15 for English↔German, IWSLT 15 for English→German, and IWSLT 14 for Turkish→English.

4.1 Data and Methods

We use Groundhog[3] as the implementation of the NMT system for all experiments (Bahdanau et al., 2015; Jean et al., 2015a). We generally follow the settings and training procedure described by Sennrich et al. (2016).

For English↔German, we report case-sensitive BLEU on detokenized text with mteval-v13a.pl for comparison to official WMT and IWSLT results. For Turkish→English, we report case-sensitive BLEU on tokenized text with multi-bleu.perl for comparison to results by Gülçehre et al. (2015).

Gülçehre et al. (2015) determine the network vocabulary based on the parallel training data,

dataset	sentences
WMT$_\text{parallel}$	4 200 000
WIT$_\text{parallel}$	200 000
WMT$_\text{mono\_de}$	160 000 000
WMT$_\text{synth\_de}$	3 600 000
WMT$_\text{mono\_en}$	118 000 000
WMT$_\text{synth\_en}$	4 200 000

Table 1: English↔German training data.

and replace out-of-vocabulary words with a special UNK symbol. They remove monolingual sentences with more than 10% UNK symbols. In contrast, we represent unseen words as sequences of subword units (Sennrich et al., 2016), and can represent any additional training data with the existing network vocabulary that was learned on the parallel data. In all experiments, the network vocabulary remains fixed.

4.1.1 English↔German

We use all parallel training data provided by WMT 2015 (Bojar et al., 2015)[4]. We use the News Crawl corpora as additional training data for the experiments with monolingual data. The amount of training data is shown in Table 1.

Baseline models are trained for a week. Ensembles are sampled from the last 4 saved models of training (saved at 12h-intervals). Each model is fine-tuned with fixed embeddings for 12 hours.

For the experiments with synthetic parallel data, we back-translate a random sample of 3 600 000 sentences from the German monolingual data set into English. The German→English system used for this is the baseline system (`parallel`). Translation took about a week on an NVIDIA Titan Black GPU. For experiments in German→English, we back-translate 4 200 000 monolingual English sentences into German, using the English→German system `+synthetic`. Note that we always use single models for back-translation, not ensembles. We leave it to future work to explore how sensitive NMT training with synthetic data is to the quality of the back-translation.

We tokenize and truecase the training data, and represent rare words via BPE (Sennrich et al., 2016). Specifically, we follow Sennrich et al. (2016) in performing BPE on the joint vocabulary with 89 500 merge operations. The network vo-

---

[3]`github.com/sebastien-j/LV_groundhog`

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

dataset	sentences
WIT	160 000
SETimes	160 000
Gigaword$_{mono}$	177 000 000
Gigaword$_{synth}$	3 200 000

Table 2: Turkish→English training data.

cabulary size is 90 000.

We also perform experiments on the IWSLT 15 test sets to investigate a cross-domain setting.[5] The test sets consist of TED talk transcripts. As in-domain training data, IWSLT provides the WIT[3] parallel corpus (Cettolo et al., 2012), which also consists of TED talks.

4.1.2 Turkish→English

We use data provided for the IWSLT 14 machine translation track (Cettolo et al., 2014), namely the WIT[3] parallel corpus (Cettolo et al., 2012), which consists of TED talks, and the SETimes corpus (Tyers and Alperen, 2010).[6] After removal of sentence pairs which contain empty lines or lines with a length ratio above 9, we retain 320 000 sentence pairs of training data. For the experiments with monolingual training data, we use the English LDC Gigaword corpus (Fifth Edition). The amount of training data is shown in Table 2. With only 320 000 sentences of parallel data available for training, this is a much lower-resourced translation setting than English↔German.

Gülçehre et al. (2015) segment the Turkish text with the morphology tool Zemberek, followed by a disambiguation of the morphological analysis (Sak et al., 2007), and removal of non-surface tokens produced by the analysis. We use the same preprocessing[7]. For both Turkish and English, we represent rare words (or morphemes in the case of Turkish) as character bigram sequences (Sennrich et al., 2016). The 20 000 most frequent words (morphemes) are left unsegmented. The networks have a vocabulary size of 23 000 symbols.

To obtain a synthetic parallel training set, we back-translate a random sample of 3 200 000 sentences from Gigaword. We use an English→Turkish NMT system trained with the same settings as the Turkish→English baseline system.

We found overfitting to be a bigger problem than with the larger English↔German data set, and follow Gülçehre et al. (2015) in using Gaussian noise (stddev 0.01) (Graves, 2011), and dropout on the output layer (p=0.5) (Hinton et al., 2012). We also use early stopping, based on BLEU measured every three hours on tst2010, which we treat as development set. For Turkish→English, we use gradient clipping with threshold 5, following Gülçehre et al. (2015), in contrast to the threshold 1 that we use for English↔German, following Jean et al. (2015a).

4.2 Results

4.2.1 English→German WMT 15

Table 3 shows English→German results with WMT training and test data. We find that mixing parallel training data with monolingual data with a dummy source side in a ratio of 1-1 improves quality by 0.4–0.5 BLEU for the single system, 1 BLEU for the ensemble. We train the system for twice as long as the baseline to provide the training algorithm with a similar amount of parallel training instances. To ensure that the quality improvement is due to the monolingual training instances, and not just increased training time, we also continued training our baseline system for another week, but saw no improvements in BLEU.

Including synthetic data during training is very effective, and yields an improvement over our baseline by 2.8–3.4 BLEU. Our best ensemble system also outperforms a syntax-based baseline (Sennrich and Haddow, 2015) by 1.2–2.1 BLEU. We also substantially outperform NMT results reported by Jean et al. (2015a) and Luong et al. (2015), who previously reported SOTA result.[8] We note that the difference is particularly large for single systems, since our ensemble is not as diverse as that of Luong et al. (2015), who used 8 independently trained ensemble components, whereas we sampled 4 ensemble components from the same training run.

4.2.2 English→German IWSLT 15

Table 4 shows English→German results on IWSLT test sets. IWSLT test sets consist of TED talks, and are thus very dissimilar from the WMT

---

[8] Luong et al. (2015) report 20.9 BLEU (tokenized) on newstest2014 with a single model, and 23.0 BLEU with an ensemble of 8 models. Our best single system achieves a *tokenized* BLEU (as opposed to untokenized scores reported in Table 3) of 23.8, and our ensemble reaches 25.0 BLEU.

name	training instances	BLEU			
		newstest2014		newstest2015	
		single	ens-4	single	ens-4
syntax-based (Sennrich and Haddow, 2015)		22.6	-	24.4	-
Neural MT (Jean et al., 2015b)		-	-	22.4	-
parallel	37m (parallel)	19.9	20.4	22.8	23.6
+monolingual	49m (parallel) / 49m (monolingual)	20.4	21.4	23.2	24.6
+synthetic	44m (parallel) / 36m (synthetic)	**22.7**	**23.8**	**25.7**	**26.5**

Table 3: English→German translation performance (BLEU) on WMT training/test sets. Ens-4: ensemble of 4 models. Number of training instances varies due to differences in training time and speed.

	name	fine-tuning		BLEU		
		data	instances	tst2013	tst2014	tst2015
	NMT (Luong and Manning, 2015) (single model)			29.4	-	-
	NMT (Luong and Manning, 2015) (ensemble of 8)			31.4	27.6	30.1
1	parallel	-	-	25.2	22.6	24.0
2	+synthetic	-	-	26.5	23.5	25.5
3	2+WIT$_{mono\_de}$	WMT$_{parallel}$ / WIT$_{mono}$	200k/200k	26.6	23.6	25.4
4	2+WIT$_{synth\_de}$	WIT$_{synth}$	200k	28.2	24.4	26.7
5	2+WIT$_{parallel}$	WIT	200k	**30.4**	**25.9**	**28.4**

Table 4: English→German translation performance (BLEU) on IWSLT test sets (TED talks). Single models.

test sets, which are news texts. We investigate if monolingual training data is especially valuable if it can be used to adapt a model to a new genre or domain, specifically adapting a system trained on WMT data to translating TED talks.

Systems 1 and 2 correspond to systems in Table 3, trained only on WMT data. System 2, trained on parallel and synthetic WMT data, obtains a BLEU score of 25.5 on tst2015. We observe that even a small amount of fine-tuning[9], i.e. continued training of an existing model, on WIT data can adapt a system trained on WMT data to the TED domain. By back-translating the monolingual WIT corpus (using a German→English system trained on WMT data, i.e. without in-domain knowledge), we obtain the synthetic data set WIT$_{synth}$. A single epoch of fine-tuning on WIT$_{synth}$ (system 4) results in a BLEU score of 26.7 on tst2015, or an improvement of 1.2 BLEU. We observed no improvement from fine-tuning on WIT$_{mono}$, the monolingual TED corpus with dummy input (system 3).

These adaptation experiments with monolingual data are slightly artificial in that parallel training data is available. System 5, which is fine-tuned with the original WIT training data, obtains a BLEU of 28.4 on tst2015, which is an improve-

---
[9]We leave the word embeddings fixed for fine-tuning.

	BLEU	
name	2014	2015
PBSMT (Haddow et al., 2015)	28.8	29.3
NMT (Gülçehre et al., 2015)	23.6	-
+shallow fusion	23.7	-
+deep fusion	24.0	-
parallel	25.9	26.7
+synthetic	**29.5**	**30.4**
+synthetic (ensemble of 4)	30.8	31.6

Table 5: German→English translation performance (BLEU) on WMT training/test sets (newstest2014; newstest2015).

ment of 2.9 BLEU. While it is unsurprising that in-domain parallel data is most valuable, we find it encouraging that NMT domain adaptation with monolingual data is also possible, and effective, since there are settings where only monolingual in-domain data is available.

The best results published on this dataset are by Luong and Manning (2015), obtained with an ensemble of 8 independently trained models. In a comparison of single-model results, we outperform their model on tst2013 by 1 BLEU.

4.2.3 German→English WMT 15

Results for German→English on the WMT 15 data sets are shown in Table 5. Like for the reverse translation direction, we see substantial improvements (3.6–3.7 BLEU) from adding monolingual training data with synthetic source sentences, which is substantially bigger than the improvement observed with deep fusion (Gülçehre et al., 2015); our ensemble outperforms the previous state of the art on newstest2015 by 2.3 BLEU.

4.2.4 Turkish→English IWSLT 14

Table 6 shows results for Turkish→English. On average, we see an improvement of 0.6 BLEU on the test sets from adding monolingual data with a dummy source side in a 1-1 ratio[10], although we note a high variance between different test sets.

With synthetic training data (Gigaword$_{synth}$), we outperform the baseline by 2.7 BLEU on average, and also outperform results obtained via shallow or deep fusion by Gülçehre et al. (2015) by 0.5 BLEU on average. To compare to what extent synthetic data has a regularization effect, even without novel training data, we also back-translate the target side of the parallel training text to obtain the training corpus parallel$_{synth}$. Mixing the original parallel corpus with parallel$_{synth}$ (ratio 1-1) gives some improvement over the baseline (1.7 BLEU on average), but the novel monolingual training data (Gigaword$_{mono}$) gives higher improvements, despite being out-of-domain in relation to the test sets. We speculate that novel in-domain monolingual data would lead to even higher improvements.

4.2.5 Back-translation Quality for Synthetic Data

One question that our previous experiments leave open is how the quality of the automatic back-translation affects training with synthetic data. To investigate this question, we back-translate the same German monolingual corpus with three different German→English systems:

- with our baseline system and greedy decoding

- with our baseline system and beam search (beam size 12). This is the same system used for the experiments in Table 3.

---

[10]We also experimented with higher ratios of monolingual data, but this led to decreased BLEU scores.

	BLEU		
	DE→EN	EN→DE	
back-translation	2015	2014	2015
none	-	20.4	23.6
parallel (greedy)	22.3	23.2	26.0
parallel (beam 12)	25.0	23.8	26.5
synthetic (beam 12)	28.3	23.9	26.6
ensemble of 3	-	24.2	27.0
ensemble of 12	-	24.7	27.6

Table 7: English→German translation performance (BLEU) on WMT training/test sets (newstest2014; newstest2015). Systems differ in how the synthetic training data is obtained. Ensembles of 4 models (unless specified otherwise).

- with the German→English system that was itself trained with synthetic data (beam size 12).

BLEU scores of the German→English systems, and of the resulting English→German systems that are trained on the different back-translations, are shown in Table 7. The quality of the German→English back-translation differs substantially, with a difference of 6 BLEU on newstest2015. Regarding the English→German systems trained on the different synthetic corpora, we find that the 6 BLEU difference in back-translation quality leads to a 0.6–0.7 BLEU difference in translation quality. This is balanced by the fact that we can increase the speed of back-translation by trading off some quality, for instance by reducing beam size, and we leave it to future research to explore how much the amount of synthetic data affects translation quality.

We also show results for an ensemble of 3 models (the best single model of each training run), and 12 models (all 4 models of each training run). Thanks to the increased diversity of the ensemble components, these ensembles outperform the ensembles of 4 models that were all sampled from the same training run, and we obtain another improvement of 0.8–1.0 BLEU.

4.3 Contrast to Phrase-based SMT

The back-translation of monolingual target data into the source language to produce synthetic parallel text has been previously explored for phrase-based SMT (Bertoldi and Federico, 2009; Lambert et al., 2011). While our approach is technically similar, synthetic parallel data fulfills novel roles

name	training		BLEU			
	data	instances	tst2011	tst2012	tst2013	tst2014
baseline (Gülçehre et al., 2015)			18.4	18.8	19.9	18.7
deep fusion (Gülçehre et al., 2015)			20.2	20.2	21.3	**20.6**
baseline	parallel	7.2m	18.6	18.2	18.4	18.3
parallel$_{synth}$	parallel/parallel$_{synth}$	6m/6m	19.9	20.4	20.1	20.0
Gigaword$_{mono}$	parallel/Gigaword$_{mono}$	7.6m/7.6m	18.8	19.6	19.4	18.2
Gigaword$_{synth}$	parallel/Gigaword$_{synth}$	8.4m/8.4m	**21.2**	**21.1**	**21.8**	20.4

Table 6: Turkish→English translation performance (*tokenized* BLEU) on IWSLT test sets (TED talks). Single models. Number of training instances varies due to early stopping.

system	BLEU	
	WMT	IWSLT
parallel	20.1	21.5
+synthetic	20.8	21.6
PBSMT gain	+0.7	+0.1
NMT gain	+2.9	+1.2

Table 8: Phrase-based SMT results (English→German) on WMT test sets (average of newstest201{4,5}), and IWSLT test sets (average of tst201{3,4,5}), and average BLEU gain from adding synthetic data for both PBSMT and NMT.
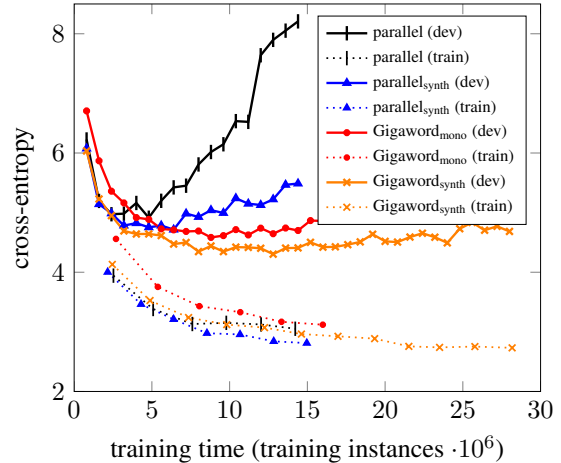

Figure 1: Turkish→English training and development set (tst2010) cross-entropy as a function of training time (number of training instances) for different systems.

in NMT.

To explore the relative effectiveness of back-translated data for phrase-based SMT and NMT, we train two phrase-based SMT systems with Moses (Koehn et al., 2007), using only WMT$_{parallel}$, or both WMT$_{parallel}$ and WMT$_{synth\_de}$ for training the translation and reordering model. Both systems contain the same language model, a 5-gram Kneser-Ney model trained on all available WMT data. We use the baseline features described by Haddow et al. (2015).

Results are shown in Table 8. In phrase-based SMT, we find that the use of back-translated training data has a moderate positive effect on the WMT test sets (+0.7 BLEU), but not on the IWSLT test sets. This is in line with the expectation that the main effect of back-translated data for phrase-based SMT is domain adaptation (Bertoldi and Federico, 2009). Both the WMT test sets and the News Crawl corpora which we used as monolingual data come from the same source, a web crawl of newspaper articles.[11] In contrast, News Crawl is out-of-domain for the IWSLT test sets.

In contrast to phrase-based SMT, which can

make use of monolingual data via the language model, NMT has so far not been able to use monolingual data to great effect, and without requiring architectural changes. We find that the effect of synthetic parallel data is not limited to domain adaptation, and that even out-of-domain synthetic data improves NMT quality, as in our evaluation on IWSLT. The fact that the synthetic data is more effective on the WMT test sets (+2.9 BLEU) than on the IWSLT test sets (+1.2 BLEU) supports the hypothesis that domain adaptation contributes to the effectiveness of adding synthetic data to NMT training.

It is an important finding that back-translated data, which is mainly effective for domain adaptation in phrase-based SMT, is more generally useful in NMT, and has positive effects that go beyond domain adaptation. In the next section, we will investigate further reasons for its effectiveness.

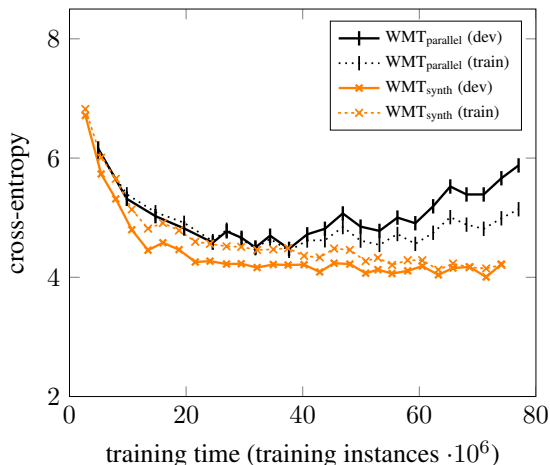[11]The WMT test sets are held-out from News Crawl.

Figure 2: English→German training and development set (newstest2013) cross-entropy as a function of training time (number of training instances) for different systems.

4.4 Analysis

We previously indicated that overfitting is a concern with our baseline system, especially on small data sets of several hundred thousand training sentences, despite the regularization employed. This overfitting is illustrated in Figure 1, which plots training and development set cross-entropy by training time for Turkish→English models. For comparability, we measure training set cross-entropy for all models on the same random sample of the parallel training set. We can see that the model trained on only parallel training data quickly overfits, while all three monolingual data sets (parallel$_{synth}$, Gigaword$_{mono}$, or Gigaword$_{synth}$) delay overfitting, and give better perplexity on the development set. The best development set cross-entropy is reached by Gigaword$_{synth}$.

Figure 2 shows cross-entropy for English→German, comparing the system trained on only parallel data and the system that includes synthetic training data. Since more training data is available for English→German, there is no indication that overfitting happens during the first 40 million training instances (or 7 days of training); while both systems obtain comparable training set cross-entropies, the system with synthetic data reaches a lower cross-entropy on the development set. One explanation for this is the domain effect discussed in the previous section.

A central theoretical expectation is that monolingual target-side data improves the model's flu-

system	produced	attested	natural
parallel	1078	53.4%	74.9%
+mono	994	61.6%	84.6%
+synthetic	1217	56.4%	82.5%

Table 9: Number of words in system output that do not occur in parallel training data (count$_{ref}$ = 1168), and proportion that is attested in data, or natural according to native speaker. English→German; newstest2015; ensemble systems.

ency, its ability to produce natural target-language sentences. As a proxy to sentence-level fluency, we investigate word-level fluency, specifically words produced as sequences of subword units, and whether NMT systems trained with additional monolingual data produce more natural words. For instance, the English→German systems translate the English phrase *civil rights protections* as a single compound, composed of three subword units: *Bürger|rechts|schutzes*[12], and we analyze how many of these multi-unit words that the translation systems produce are well-formed German words.

We compare the number of words in the system output for the newstest2015 test set which are produced via subword units, and that do not occur in the parallel training corpus. We also count how many of them are attested in the full monolingual corpus or the reference translation, which we all consider 'natural'. Additionally, the main authors, a native speaker of German, annotated a random subset ($n = 100$) of unattested words of each system according to their naturalness[13], distinguishing between natural German words (or names) such as *Literatur|klassen* 'literature classes', and nonsensical ones such as **As|best|atten* (a misspelling of *Astbestmatten* 'asbestos mats').

In the results (Table 9), we see that the systems trained with additional monolingual or synthetic data have a higher proportion of novel words attested in the non-parallel data, and a higher proportion that is deemed natural by our annotator. This supports our expectation that additional monolingual data improves the (word-level) fluency of the NMT system.

---

[12]Subword boundaries are marked with '|'.
[13]For the annotation, the words were blinded regarding the system that produced them.

5 Related Work

To our knowledge, the integration of monolingual data for pure neural machine translation architectures was first investigated by (Gülçehre et al., 2015), who train monolingual language models independently, and then integrate them during decoding through rescoring of the beam (*shallow fusion*), or by adding the recurrent hidden state of the language model to the decoder state of the encoder-decoder network, with an additional controller mechanism that controls the magnitude of the LM signal (*deep fusion*). In deep fusion, the controller parameters and output parameters are tuned on further parallel training data, but the language model parameters are fixed during the fine-tuning stage. Jean et al. (2015b) also report on experiments with reranking of NMT output with a 5-gram language model, but improvements are small (between 0.1–0.5 BLEU).

The production of synthetic parallel texts bears resemblance to data augmentation techniques used in computer vision, where datasets are often augmented with rotated, scaled, or otherwise distorted variants of the (limited) training set (Rowley et al., 1996).

Another similar avenue of research is self-training (McClosky et al., 2006; Schwenk, 2008). The main difference is that self-training typically refers to scenario where the training set is enhanced with training instances with artificially produced output labels, whereas we start with human-produced output (i.e. the translation), and artificially produce an input. We expect that this is more robust towards noise in the automatic translation. Improving NMT with monolingual source data, following similar work on phrase-based SMT (Schwenk, 2008), remains possible future work.

Domain adaptation of neural networks via continued training has been shown to be effective for neural language models by (Ter-Sarkisov et al., 2015), and in work parallel to ours, for neural translation models (Luong and Manning, 2015). We are the first to show that we can effectively adapt neural translation models with monolingual data.

6 Conclusion

In this paper, we propose two simple methods to use monolingual training data during training of NMT systems, with no changes to the network architecture. Providing training examples with dummy source context was successful to some extent, but we achieve substantial gains in all tasks, and new SOTA results, via back-translation of monolingual target data into the source language, and treating this synthetic data as additional training data. We also show that small amounts of in-domain monolingual data, back-translated into the source language, can be effectively used for domain adaptation. In our analysis, we identified domain adaptation effects, a reduction of overfitting, and improved fluency as reasons for the effectiveness of using monolingual data for training.

While our experiments did make use of monolingual training data, we only used a small random sample of the available data, especially for the experiments with synthetic parallel data. It is conceivable that larger synthetic data sets, or data sets obtained via data selection, will provide bigger performance benefits.

Because we do not change the neural network architecture to integrate monolingual training data, our approach can be easily applied to other NMT systems. We expect that the effectiveness of our approach not only varies with the quality of the MT system used for back-translation, but also depends on the amount (and similarity to the test set) of available parallel and monolingual data, and the extent of overfitting of the baseline model. Future work will explore the effectiveness of our approach in more settings.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*

StatMT 09. Association for Computational Linguistics.

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

P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014. In *Proceedings of the 11th Workshop on Spoken Language Translation*, pages 2–16, Lake Tahoe, CA, USA.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. 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)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Alex Graves. 2011. Practical Variational Inference for Neural Networks. In J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2348–2356. Curran Associates, Inc.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On Using Monolingual Corpora in Neural Machine Translation. *CoRR*, abs/1503.03535.

Barry Haddow, Matthias Huck, Alexandra Birch, Nikolay Bogoychev, and Philipp Koehn. 2015. The Edinburgh/JHU Phrase-based Machine Translation Systems for WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 126–133, Lisbon, Portugal. Association for Computational Linguistics.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015a. On Using Very Large Target Vocabulary for Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Beijing, China. Association for Computational Linguistics.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015b. Montreal Neural Machine Translation Systems for WMT'15 . In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL-2007 Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on Translation Model Adaptation Using Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 284–293, Edinburgh, Scotland. Association for Computational Linguistics.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford Neural Machine Translation Systems for Spoken Language Domains. In *Proceedings of the International Workshop on Spoken Language Translation 2015*, Da Nang, Vietnam.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. 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. Association for Computational Linguistics.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective Self-training for Parsing. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 152–159, New York. Association for Computational Linguistics.

Henry Rowley, Shumeet Baluja, and Takeo Kanade. 1996. Neural Network-Based Face Detection. In *Computer Vision and Pattern Recognition '96*.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2007. Morphological Disambiguation of Turkish Text with Perceptron Algorithm. In *CICLing 2007*, pages 107–118.

Holger Schwenk. 2008. Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. In *International Workshop on Spoken Language Translation*, pages 182–189.

Rico Sennrich and Barry Haddow. 2015. A Joint Dependency Model of Morphological and Syntactic Structure for Statistical Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2081–2087, Lisbon, Portugal. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin, Germany.

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

Alex Ter-Sarkisov, Holger Schwenk, Fethi Bougares, and Loïc Barrault. 2015. Incremental Adaptation Strategies for Neural Network Language Models. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 48–56, Beijing, China. Association for Computational Linguistics.

Francis M. Tyers and Murat S. Alperen. 2010. SE-Times: A parallel corpus of Balkan languages. In *Workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages at the Language Resources and Evaluation Conference*, pages 1–5.