Tagged Back-translation Revisited: Why Does It Really Work?

Benjamin Marie **Raphael Rubino** **Atsushi Fujita**
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{bmarie, raphael.rubino, atsushi.fujita}@nict.go.jp

Abstract

In this paper, we show that neural machine translation (NMT) systems trained on large back-translated data overfit some of the characteristics of machine-translated texts. Such NMT systems better translate human-produced translations, i.e., translationese, but may largely worsen the translation quality of original texts. Our analysis reveals that adding a simple tag to back-translations prevents this quality degradation and improves on average the overall translation quality by helping the NMT system to distinguish back-translated data from original parallel data during training. We also show that, in contrast to high-resource configurations, NMT systems trained in low-resource settings are much less vulnerable to overfit back-translations. We conclude that the back-translations in the training data should always be tagged especially when the origin of the text to be translated is unknown.

1 Introduction

During training, neural machine translation (NMT) can leverage a large amount of monolingual data in the target language. Among existing ways of exploiting monolingual data in NMT, the so-called *back-translation* of monolingual data (Sennrich et al., 2016a) is undoubtedly the most prevalent one, as it remains widely used in state-of-the-art NMT systems (Barrault et al., 2019). NMT systems trained on back-translated data can generate more fluent translations (Sennrich et al., 2016a) thanks to the use of much larger data in the target language to better train the decoder, especially for low-resource conditions where only a small quantity of parallel training data is available. However, the impact of the noisiness of the synthetic source sentences generated by NMT largely remains unclear and understudied. Edunov et al. (2018) even showed that introducing synthetic noise in back-translations actually improves translation quality and enables the use of a much larger quantity of back-translated data for further improvements in translation quality. More recently, Caswell et al. (2019) empirically demonstrated that adding a unique token at the beginning of each back-translation acts as a tag that helps the system during training to differentiate back-translated data from the original parallel training data and is as effective as introducing synthetic noise for improving translation quality. It is also much simpler since it requires only one editing operation, adding the tag, and non-parametric. However, it is not fully understood why adding a tag has such a significant impact and to what extent it helps to distinguish back-translated data from the original parallel data.

In this paper, we report on the impact of tagging back-translations in NMT, focusing on the following research questions (see Section 2 for our motivation).

Q1. Do NMT systems trained on large back-translated data capture some of the characteristics of human-produced translations, i.e., *translationese*?

Q2. Does a tag for back-translations really help differentiate translationese from original texts?

Q3. Are NMT systems trained on back-translation for low-resource conditions as sensitive to translationese as in high-resource conditions?

2 Motivation

During the training with back-translated data (Sennrich et al., 2016a), we can expect the NMT system to learn the characteristics of back-translations, i.e., translations generated by NMT, and such characteristics will be consequently exhibited at test time. However, translating translations is a rather artificial task, whereas users usually want to perform translation of original texts. Nonetheless, many

of the test sets used by the research community for evaluating MT systems actually contain a large portion of texts that are translations produced by humans, i.e., *translationese*. Translationese texts are known to be much simpler, with a lower mean sentence length and more standardized than original texts (Laviosa-Braithwaite, 1998). These characteristics overlap with those of translations generated by NMT systems that have been shown simpler, shorter, and to exhibit a less diverse vocabulary than original texts (Burlot and Yvon, 2018). These similarities raise **Q1**.

Caswell et al. (2019) hypothesized that tagging back-translations helps the NMT system during training to make some distinction between the back-translated data and the original parallel data. Even though the effectiveness of a tag has been empirically demonstrated, the nature of this distinction remains unclear. Thus, we pose **Q2**.

The initial motivation for back-translation is to improve NMT for low-resource language pairs by augmenting the training data. Therefore, we verify whether our answers to Q1 and Q2 for high-resource conditions are also valid in low-resource conditions, answering **Q3**.

3 Experiments

3.1 Data

As parallel data for training our NMT systems, we used all the parallel data provided for the shared translation tasks of WMT19[1] for English–German (en-de), excluding the Paracrawl corpus, and WMT15[2] for English–French (en-fr).[3] As monolingual data for each of English, German, and French to be used for back-translation, we concatenated all the News Crawl corpora provided by WMT, and randomly extracted 25M sentences. For our simulation of low-resource conditions, we randomly sub-sampled 200k sentence pairs from the parallel data to train NMT systems and used these systems to back-translate 1M sentences randomly sub-sampled from the monolingual data. For validation, i.e., selecting the best model after training, we chose newstest2016 for en-de and newstest2013 for en-fr, since they are rather balanced on their source side between translationese and original texts. For evaluation, since most of the WMT test sets are made of both original and translationese texts, we used all the newstest sets, from WMT10 to WMT19 for en-de, and from WMT08 to WMT15 for en-fr.[4]

All our data were pre-processed in the same way: we performed tokenization and truecasing with `Moses` (Koehn et al., 2007).

3.2 NMT Systems

For NMT, we used the Transformer (Vaswani et al., 2017) implemented in Marian (Junczys-Dowmunt et al., 2018) with standard hyper-parameters for training a Transformer base model.[5] To compress the vocabulary, we learned 32k byte-pair encoding (BPE) operations (Sennrich et al., 2016b) for each side of the parallel training data.

The back-translations were generated through decoding with Marian the sampled monolingual sentences using beam search with a beam size of 12 and a length normalization of 1.0. The back-translated data were then concatenated to the original parallel data and a new NMT model was trained from scratch using the same hyper-parameters used to train the model that generated the back-translations.

We evaluated all systems with BLEU (Papineni et al., 2002) computed by sacreBLEU (Post, 2018). To evaluate only on the part of the test set that have original text or translationese on the source side, we used the `--origlang` option of sacreBLEU with the value "non-L1" for translationese texts and "L1" for original texts, where L1 is the source language, and report on their respective BLEU scores.[6]

3.3 Results in Resource-Rich Conditions

Our results with back-translations (BT) and tagged back-translations (T-BT) are presented in Table 1. When using BT, we consistently observed a drop of BLEU scores for original texts for all the translations tasks, with the largest drop of 12.1 BLEU points (en→fr, 2014). Conversely, BLEU scores for translationese texts were improved for most tasks, with the largest gain of 10.4 BLEU points

[1] http://www.statmt.org/wmt19/translation-task.html

[2] http://www.statmt.org/wmt15/translation-task.html

[3] After pre-processing and cleaning, we obtained 5.2M and 32.8M sentence pairs for en-de and en-fr, respectively.

[4] For WMT14, we used the "full" version instead of the default filtered version in sacreBLEU that does not contain information on the origin of the source sentences.

[5] The full list of hyper-parameters is provided in the supplementary material (Appendix A).

[6] sacreBLEU signatures where "L1" and "L2" respectively indicates a two-letter identifier for the source and target languages of either de-en, en-de, fr-en, or en-fr, and "XXX" the name of the test set: BLEU+case.mixed+lang.L1-L2+numrefs.1+{origlang.L1,origlang.non-L2}+smooth.exp+test.XXX+tok.13a+version.1.4.2

System	test set	de→en all	de→en o	de→en n-o	en→de all	en→de o	en→de n-o
BT	2010	28.9 (+0.5)	33.2 (-0.9)	27.9 (+0.7)	21.8 (-2.3)	24.6 (-5.7)	21.0 (-1.2)
	2011	25.3 (-0.3)	29.9 (-1.0)	24.2 (-0.2)	19.9 (-1.4)	23.8 (-1.9)	19.0 (-1.1)
	2012	27.1 (+0.3)	27.9 (-1.6)	27.0 (+0.7)	20.4 (-1.2)	24.5 (-4.6)	19.3 (-0.2)
	2013	30.3 (+0.3)	34.7 (-1.6)	29.2 (+0.6)	23.8 (-1.9)	25.1 (-2.8)	23.6 (-1.7)
	2014	32.8 (+2.2)	27.4 (-2.5)	36.8 (+7.0)	25.4 (-0.5)	23.2 (-3.3)	27.9 (+2.7)
	2015	33.8 (+2.4)	22.5 (-1.9)	39.5 (+5.5)	27.2 (-1.1)	28.1 (-2.9)	24.7 (+1.9)
	2017	35.5 (+3.0)	27.2 (-1.1)	42.8 (+7.4)	26.4 (-0.1)	26.3 (-3.6)	25.5 (+3.3)
	2018	43.9 (+4.6)	32.0 (-1.0)	53.8 (+10.4)	38.0 (-1.4)	38.9 (-5.9)	35.0 (+3.8)
	2019	-	33.1 (-1.5)	-	-	31.4 (-4.8)	-
T-BT	2010	29.5 (+1.1)	34.4 (+0.3)	28.4 (+1.2)	25.0 (+0.9)	30.5 (+0.2)	23.4 (+1.2)
	2011	26.4 (+0.8)	31.7 (+0.8)	25.2 (+0.8)	22.1 (+0.8)	25.8 (+0.1)	21.0 (+0.9)
	2012	28.1 (+1.3)	30.2 (+0.7)	27.7 (+1.4)	22.8 (+1.2)	30.0 (+0.9)	20.9 (+1.4)
	2013	30.8 (+0.8)	36.0 (-0.3)	29.6 (+1.0)	26.4 (+0.7)	28.1 (+0.2)	26.1 (+0.8)
	2014	32.4 (+1.8)	29.6 (-0.3)	33.8 (+4.0)	27.9 (+2.0)	26.7 (+0.2)	29.4 (+4.2)
	2015	33.9 (+2.5)	24.9 (+0.5)	37.7 (+3.7)	29.9 (+1.6)	32.1 (+1.1)	25.6 (+2.8)
	2017	35.5 (+3.0)	28.1 (-0.2)	41.2 (+5.8)	28.7 (+2.2)	30.7 (+0.8)	26.0 (+3.8)
	2018	43.2 (+3.9)	33.0 (+0.0)	50.4 (+7.0)	41.8 (+2.4)	45.6 (+0.8)	35.5 (+4.3)
	2019	-	35.0 (+0.4)	-	-	37.6 (+1.4)	-

System	test set	fr→en all	fr→en o	fr→en n-o	en→fr all	en→fr o	en→fr n-o
BT	2008	22.9 (-1.7)	27.9 (-2.6)	22.2 (-1.5)	23.2 (-0.2)	21.2 (-3.3)	23.6 (+0.5)
	2009	26.5 (-2.3)	41.1 (-5.3)	23.9 (-1.6)	27.7 (+1.1)	22.7 (-2.0)	28.4 (+1.4)
	2010	29.3 (-1.4)	27.4 (-7.8)	29.5 (+0.5)	28.2 (-0.5)	22.5 (-11.1)	29.8 (+2.5)
	2011	29.4 (-1.9)	29.3 (-4.7)	29.4 (-1.1)	30.9 (+0.0)	36.7 (-8.2)	29.3 (+2.1)
	2012	29.7 (-1.4)	34.3 (-4.3)	28.6 (-0.6)	28.4 (+1.1)	26.3 (-4.1)	29.0 (+2.5)
	2014	36.6 (+0.6)	31.4 (-4.7)	40.3 (+5.6)	32.9 (-3.1)	26.1 (-12.1)	39.6 (+6.1)
	2015	36.2 (+0.0)	40.9 (-3.1)	29.8 (+3.5)	35.7 (+1.7)	25.1 (-4.4)	44.9 (+6.5)
T-BT	2008	24.5 (-0.1)	29.5 (-1.0)	23.7 (+0.0)	23.8 (+0.4)	25.1 (+0.6)	23.5 (+0.4)
	2009	28.9 (+0.1)	46.4 (+0.0)	25.7 (+0.2)	27.3 (+0.7)	25.1 (+0.4)	27.7 (+0.7)
	2010	31.2 (+0.5)	35.1 (-0.1)	29.6 (+0.6)	30.0 (+1.3)	34.1 (+0.5)	28.9 (+1.6)
	2011	31.8 (+0.5)	33.3 (-0.7)	31.4 (+0.9)	31.6 (+0.7)	45.3 (+0.4)	28.0 (+0.8)
	2012	31.8 (+0.7)	38.3 (-0.3)	30.1 (+0.9)	28.9 (+1.6)	31.9 (+1.5)	28.1 (+1.6)
	2014	37.3 (+1.3)	36.1 (+0.0)	37.2 (+2.5)	38.2 (+2.2)	39.7 (+1.5)	36.5 (+3.0)
	2015	36.6 (+0.4)	43.2 (-0.8)	27.9 (+1.6)	36.0 (+2.0)	30.7 (+1.2)	41.2 (+2.8)

Table 1: BLEU scores for NMT systems trained with back-translations (BT) and tagged back-translations (T-BT) for each origin of the source text: original (o) or translationese (n-o). The values in parentheses are the differences between the BLEU scores of the evaluated system and the vanilla system trained without any back-translated data.

(de→en, 2018). These results give an answer to **Q1**: NMT overfits back-translations, potentially due to their much larger size than the original parallel data used for training. Interestingly, using back-translations does not consistently improve translation quality. We assume that newstest sets may manifest some different characteristics of translationese from one year to another.

Prepending a tag (T-BT) had a strong impact on the translation quality for original texts, recovering or even surpassing the quality obtained by the NMT system without back-translated data, always beating BT. The large improvements of BLEU scores over BT show that a tag helps in identifying translationese (answer for **Q2**). In the supplementary material (Appendix B), we present additional results obtained using more back-translations (up to 150M sentences) showing a similar impact of tags. However, while a tag in such a configuration prevents an even larger drop of the BLEU scores, it is not sufficient to attain a BLEU score similar to the configurations that use less back-translations.

Interestingly, the best NMT system was not always the same depending on the translation direction and the origin of the test sets. It is thus possible to select either of the models to obtain the best translation quality given the origin of the source sentences, according to the results on the validation set for instance.[7]

[7]Since this observation is rather secondary, we present results for best model selection in the supplementary material (Appendix C). Note also that these BLEU scores can potentially be further increased by using a validation set whose source side is either original texts or translationese respectively to translate original texts or translationese at test time.

System	test set	de→en			en→de		
		all	o	n-o	all	o	n-o
BT	2010	24.1 (+9.5)	27.1 (+12.4)	23.3 (+8.8)	18.0 (+2.9)	21.6 (+2.7)	17.0 (+3.0)
	2011	21.0 (+8.1)	23.9 (+10.3)	20.3 (+7.6)	16.3 (+2.3)	19.1 (+2.9)	15.6 (+2.1)
	2012	22.2 (+8.6)	21.6 (+8.7)	22.3 (+8.5)	16.4 (+2.5)	19.8 (+2.6)	15.5 (+2.5)
	2013	25.0 (+9.0)	28.1 (+9.6)	24.1 (+8.7)	19.6 (+2.9)	20.0 (+3.2)	19.5 (+2.8)
	2014	25.1 (+11.3)	20.9 (+8.4)	27.7 (+13.3)	19.7 (+4.5)	18.7 (+3.3)	20.3 (+6.1)
	2015	27.1 (+11.8)	18.4 (+6.9)	31.0 (+14.3)	21.5 (+4.0)	22.5 (+3.6)	18.3 (+5.0)
	2017	27.6 (+12.5)	21.5 (+8.2)	32.4 (+16.2)	20.7 (+4.0)	20.8 (+2.7)	19.3 (+5.5)
	2018	34.3 (+16.4)	25.2 (+10.7)	41.0 (+21.1)	29.3 (+6.7)	30.4 (+5.4)	26.3 (+8.3)
	2019	-	26.1 (+11.9)	-	-	24.8 (+4.8)	-
T-BT	2010	24.4 (+9.8)	27.4 (+12.7)	23.6 (+9.1)	18.8 (+3.7)	22.6 (+3.7)	17.7 (+3.7)
	2011	21.8 (+8.9)	25.3 (+11.7)	20.9 (+8.2)	16.8 (+2.8)	20.2 (+4.0)	16.0 (+2.5)
	2012	22.8 (+9.2)	22.9 (+10.0)	22.8 (+9.0)	17.2 (+3.3)	21.3 (+4.1)	16.1 (+3.1)
	2013	25.9 (+9.9)	29.4 (+10.9)	24.9 (+9.5)	20.2 (+3.5)	20.5 (+3.7)	20.2 (+3.5)
	2014	25.1 (+11.3)	22.1 (+9.6)	26.8 (+12.4)	20.1 (+4.9)	19.5 (+4.1)	20.6 (+6.4)
	2015	27.0 (+11.7)	19.4 (+7.9)	30.5 (+13.8)	22.0 (+4.5)	23.5 (+4.6)	18.2 (+4.9)
	2017	27.8 (+12.7)	22.5 (+9.2)	32.0 (+15.8)	21.1 (+4.4)	22.2 (+4.1)	19.2 (+5.4)
	2018	34.2 (+16.3)	26.4 (+11.9)	39.8 (+19.9)	30.5 (+7.9)	32.9 (+7.9)	25.5 (+7.5)
	2019	-	26.8 (+12.6)	-	-	26.9 (+6.9)	-

System	test set	fr→en			en→fr		
		all	o	n-o	all	o	n-o
BT	2008	20.5 (+2.8)	26.3 (+1.6)	19.7 (+3.0)	21.3 (+4.1)	21.2 (+3.2)	21.3 (+4.3)
	2009	24.0 (+3.3)	39.7 (+5.4)	21.2 (+3.1)	24.8 (+6.1)	21.6 (+5.1)	25.2 (+6.2)
	2010	26.4 (+4.7)	28.3 (+4.0)	25.4 (+5.0)	26.1 (+6.0)	29.9 (+6.3)	24.9 (+5.8)
	2011	26.2 (+3.4)	26.9 (+0.7)	26.0 (+4.1)	28.1 (+6.3)	38.1 (+8.6)	25.5 (+5.8)
	2012	26.4 (+4.0)	31.4 (+1.3)	25.2 (+4.6)	26.4 (+6.3)	27.2 (+5.9)	26.1 (+6.3)
	2014	32.2 (+7.8)	28.9 (+4.5)	33.6 (+10.5)	31.4 (+7.6)	28.9 (+4.5)	32.9 (+10.3)
	2015	30.0 (+5.9)	34.0 (+5.0)	24.8 (+7.1)	29.9 (+8.0)	23.7 (+5.4)	35.5 (+10.1)
T-BT	2008	21.3 (+3.6)	27.5 (+2.8)	20.4 (+3.7)	20.8 (+3.6)	21.7 (+3.7)	20.6 (+3.6)
	2009	24.6 (+3.9)	41.6 (+7.3)	21.5 (+3.4)	23.7 (+5.0)	20.8 (+4.3)	24.1 (+5.1)
	2010	27.0 (+5.3)	29.6 (+5.3)	25.7 (+5.3)	25.6 (+5.5)	29.8 (+6.2)	24.3 (+5.2)
	2011	27.4 (+4.6)	29.7 (+3.5)	26.7 (+4.8)	27.3 (+5.5)	36.9 (+7.4)	24.8 (+5.1)
	2012	27.3 (+4.9)	33.3 (+3.2)	25.7 (+5.1)	25.6 (+5.5)	26.8 (+5.5)	25.2 (+5.4)
	2014	31.8 (+7.4)	29.9 (+5.5)	32.1 (+9.0)	31.0 (+7.2)	30.4 (+6.0)	30.9 (+8.3)
	2015	30.6 (+6.5)	35.6 (+6.6)	23.7 (+6.0)	29.2 (+7.3)	24.0 (+5.7)	34.2 (+8.8)

Table 2: BLEU scores for low-resource configurations.

3.4 Results in Low-Resource Conditions

In low-resource conditions, as reported in Table 2, the translation quality can be notably improved by adding back-translations. Using BT, we observed improvements of BLEU scores ranging from 0.7 (fr→en, 2011) to 12.4 (de→en, 2010) BLEU points for original texts and from 2.1 (en→de, 2011) to 21.1 (de→en, 2018) BLEU points for translationese texts. These results remain in line with one of the initial motivations for using back-translation: improving translation quality in low-resource conditions. In this setting without back-translated data, the data in the target language is too small for the NMT system to learn reasonably good representations for the target language. Adding 5 times more data in the target language, through back-translation, clearly helps the systems without any negative impact of the noisiness of the back-translations that were generated by the initial system. We assume here that since the quality of the back-translations is very low, their characteristics are quite different from the ones of translationese texts. This is confirmed by our observation that adding the tag has only a negligible impact on the BLEU scores for all the tasks (answer to **Q3**).

3.5 Tagged Test Sets

A tag on back-translations helps identifying translationese during NMT training. Thus, adding the same tag on the test sets should have a very different impact depending on the origin of the source sentences. If we tag original sentences and decode them with a T-BT model, then we enforce the decoding of translationese. Since we mislead the decoder, translation quality should drop. On the other hand, by tagging translationese sentences, we help the decoder that can now rely on the tag to be very confident that the text to decode is translationese.

Our results presented in Table 3 confirm these

System	de→en		en→de		fr→en		en→fr	
	2017	2018	2017	2018	2012	2015	2012	2015
tagged original	-2.0	-2.6	-5.9	-9.6	-7.5	-4.9	-10.1	-11.1
tagged non-original	+1.6	+3.4	+0.8	+1.6	-3.1	+1.4	-0.3	+3.6

Table 3: Results with tagged test sets, either original or non-original, decoded with the T-BT model in the high-resource condition. Delta BLEU scores are computed relatively to the configurations with untagged test sets.

assumptions. We observed a drop of BLEU scores when decoding tagged original texts with the T-BT model, while we saw an improvement of translation quality for 6 out of 8 test sets when decoding tagged translationese texts. The remaining 2 test sets for which we did not observed any improvements are newstest2012 for both translation directions of en-fr. It potentially indicates a mismatch between the characteristics of translationese in newstest2012 and those exhibited by back-translations used for training the T-BT model.

4 Discussions

We empirically demonstrated that training NMT on back-translated data overfits some of its characteristics that are partly similar to those of translationese. Using back-translation improves translation quality for translationese texts but worsens it for original texts. Previous work (Graham et al., 2019; Zhang and Toral, 2019) showed that state-of-the-art NMT systems are better in translating translationese than original texts. Our results show that this is partly due to the use of back-translations which is also confirmed by concurrent and independent work (Bogoychev and Sennrich, 2019; Edunov et al., 2019). Adding a tag to back-translations prevents a large drop of translation quality on original texts while improvements of translation quality for translationese texts remain and may be further boosted by tagging test sentences at decoding time. Moreover, in low-resource conditions, we show that the overall tendency is significantly different from the high-resource conditions: back-translation improves translation quality for both translationese and original texts while adding a tag to back-translations has only a little impact.

We conclude from this study that training NMT on back-translated data, in high-resource conditions, remains reasonable when the user knows in advance that the system will be used to translate translationese texts. If the user does not know it a priori, a tag should be added to back-translations during training to prevent a possible large drop of translation quality.

For future work, following the work on automatic identification of translationese (Rabinovich and Wintner, 2015; Rubino et al., 2016), we plan to investigate the impact of tagging translationese texts inside parallel training data, such as parallel sentences collected from the Web.

Acknowledgments

We would like to thank the reviewers for their useful comments and suggestions. A part of this work was conducted under the program "Research and Development of Enhanced Multilingual and Multipurpose Speech Translation System" of the Ministry of Internal Affairs and Communications (MIC), Japan. Benjamin Marie was partly supported by JSPS KAKENHI Grant Number 20K19879 and the tenure-track researcher start-up fund in NICT. Atsushi Fujita was partly supported by JSPS KAKENHI Grant Number 19H05660.

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, translationese and noise in synthetic data for neural machine translation. *arXiv preprint arXiv:1911.03362*.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2019. On the evaluation of machine translation systems trained with back-translation. *arXiv preprint arXiv:1908.05204*.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *arXiv preprint arXiv:1906.09833*.

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++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. 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 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Sara Laviosa-Braithwaite. 1998. Universals of translation. *Routledge encyclopedia of translation studies. London: Routledge*, pages 288–291.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–970, San Diego, USA. Association for Computational Linguistics.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

A NMT system hyper-parameters

For training NMT systems with `Marian` 1.7.6 (1d4ba73), we used the hyper-parameters, on 8 GPUs, presented by Table 4 and kept the remaining ones with their default values.

```
--type transformer
--train-sets para.L1 para.L2
--model model.npz --max-length
150 --mini-batch-fit
--valid-freq 5000 --save-freq
5000 --workspace 4000
--disp-freq 500 --valid-sets
dev.bpe32k.L1 dev.bpe32k.L2
--beam-size 12 --normalize=1
--valid-mini-batch 16
--overwrite --early-stopping
5 --cost-type=ce-mean-words
--valid-metrics ce-mean-words
bleu --keep-best
--enc-depth 6 --dec-depth
6 --transformer-dropout
0.1 --learn-rate 0.0003
--lr-warmup 16000
--lr-decay-inv-sqrt 16000
--label-smoothing 0.1
--dim-vocabs 32000 32000
--optimizer-params 0.9 0.98
1e-09 --clip-norm 5 --sync-sgd
--exponential-smoothing
```

Table 4: Parameters of Marian used for training our NMT systems.

B Experiments with Larger Quantity of Back-translaitons

Table 5 presents the results using much larger back-translations in the high-resource conditions.

C Best Model Selection

As discussed in Section 3.3, among the original model, the one trained with back-translation (BT), and the one trained with tagged back-translation (T-BT), the best-performing model is not always the same depending on the translation direction. For de→en and en→de, the best model is always T-BT. However, for fr→en, the system that does not use any back-translation is the best to translate original texts while T-BT is the best for translationese texts. For en→fr, the best system for translating translationese texts is BT while the best system for translating original texts is T-BT. This selection is performed by evaluating the translation quality for each model on the validation sets original and translationese texts.

By applying this selection strategy, we can significantly improve the overall translation quality for given test sets, as reported in Table 6.

System	test set	de→en all	de→en o	de→en n-o	en→de all	en→de o	en→de n-o
BT	2010	28.7 (+0.3)	32.0 (-2.1)	27.9 (+0.7)	22.3 (-1.8)	25.8 (-4.5)	21.3 (-0.9)
	2011	24.6 (-1.0)	29.2 (-1.7)	23.5 (-0.9)	19.9 (-1.4)	23.1 (-2.6)	19.1 (-1.0)
	2012	26.4 (-0.4)	27.1 (-2.4)	26.2 (-0.1)	20.7 (-0.9)	25.2 (-3.9)	19.5 (+0.0)
	2013	29.6 (-0.4)	33.1 (-3.2)	28.6 (+0.0)	23.8 (-1.9)	24.4 (-3.5)	23.7 (-1.6)
	2014	32.4 (+1.8)	25.7 (-4.2)	37.3 (+7.5)	26.0 (+0.1)	23.4 (-3.1)	28.9 (+3.7)
	2015	33.4 (+2.0)	21.2 (-3.2)	39.4 (+5.4)	27.4 (-0.9)	27.7 (-3.3)	25.7 (+2.9)
	2017	34.6 (+2.1)	25.7 (-2.6)	42.2 (+6.8)	26.6 (+0.1)	25.9 (-4.0)	26.4 (+4.2)
	2018	43.2 (+3.9)	30.1 (-2.9)	53.9 (+10.5)	38.1 (-1.3)	38.8 (-6.0)	35.4 (+4.2)
	2019	-	31.4 (-3.2)	-	-	32.1 (-4.1)	-
T-BT	2010	29.5 (+1.1)	34.1 (+0.0)	28.3 (+1.1)	24.9 (+0.8)	29.3 (-1.0)	23.7 (+1.5)
	2011	25.9 (+0.3)	30.4 (-0.5)	24.8 (+0.4)	21.9 (+0.6)	26.0 (+0.3)	20.7 (+0.6)
	2012	27.5 (+0.7)	28.8 (-0.7)	27.3 (+1.0)	22.7 (+1.1)	28.8 (-0.3)	21.1 (+1.6)
	2013	30.7 (+0.7)	35.2 (-1.1)	29.6 (+1.0)	26.1 (+0.4)	27.4 (-0.5)	25.9 (+0.6)
	2014	32.5 (+1.9)	28.2 (-1.7)	35.4 (+5.6)	28.2 (+2.3)	26.8 (+0.3)	30.0 (+4.8)
	2015	33.7 (+2.3)	23.7 (-0.7)	38.3 (+4.3)	29.6 (+1.3)	31.1 (+0.1)	26.7 (+3.9)
	2017	35.2 (+2.7)	27.3 (-1.0)	41.5 (+6.1)	28.3 (+1.8)	29.8 (-0.1)	26.3 (+4.1)
	2018	43.4 (+4.1)	32.4 (-0.6)	51.5 (+8.1)	41.7 (+2.3)	45.0 (+0.2)	36.1 (+4.9)
	2019	-	34.3 (-0.3)	-	-	36.5 (+0.3)	-

System	test set	fr→en all	fr→en o	fr→en n-o	en→fr all	en→fr o	en→fr n-o
BT	2008	20.8 (-3.8)	27.4 (-3.1)	19.8 (-3.9)	21.6 (-1.8)	17.5 (-7.0)	22.5 (-0.6)
	2009	23.9 (-4.9)	38.3 (-8.1)	21.2 (-4.3)	26.4 (-0.2)	20.4 (-4.3)	27.3 (+0.3)
	2010	27.2 (-3.5)	27.7 (-7.5)	27.0 (-2.0)	26.7 (-2.0)	19.2 (-14.4)	28.8 (+1.5)
	2011	27.3 (-4.0)	27.3 (-6.7)	27.3 (-3.2)	28.9 (-2.0)	31.4 (-13.5)	28.2 (+1.0)
	2012	26.8 (-4.3)	31.4 (-7.2)	25.7 (-3.5)	26.5 (-0.8)	22.2 (-8.2)	27.7 (+1.2)
	2014	33.5 (-2.5)	28.8 (-7.3)	36.9 (+2.2)	29.9 (-6.1)	20.6 (-17.6)	39.4 (+5.9)
	2015	31.7 (-4.5)	35.5 (-8.5)	27.1 (+0.8)	32.4 (-1.6)	18.4 (-11.1)	44.9 (+6.5)
T-BT	2008	24.7 (+0.1)	30.6 (+0.1)	23.8 (+0.1)	24.1 (+0.7)	25.6 (+1.1)	23.7 (+0.6)
	2009	28.4 (-0.4)	45.3 (-1.1)	25.2 (-0.3)	27.7 (+1.1)	25.7 (+1.0)	28.0 (+1.0)
	2010	31.2 (+0.5)	34.2 (-1.0)	29.8 (+0.8)	30.6 (+1.9)	34.5 (+0.9)	29.5 (+2.2)
	2011	31.8 (+0.5)	32.7 (-1.3)	31.5 (+1.0)	31.6 (+0.7)	45.5 (+0.6)	28.0 (+0.8)
	2012	31.6 (+0.5)	37.5 (-1.1)	30.2 (+1.0)	29.2 (+1.9)	31.9 (+1.5)	28.4 (+1.9)
	2014	37.9 (+1.9)	35.6 (-0.5)	38.7 (+4.0)	38.5 (+2.5)	39.7 (+1.5)	37.0 (+3.5)
	2015	36.2 (+0.0)	42.2 (-1.8)	28.3 (+2.0)	36.3 (+2.3)	30.2 (+0.7)	42.1 (+3.7)

Table 5: BLEU scores for all the systems in the high-resource conditions using 150M back-translations or the entire news crawl corpus for en→fr (76.6M sentences).

Sys.	2008 fr→en	2008 en→fr	2009 fr→en	2009 en→fr	2010 fr→en	2010 en→fr	2011 fr→en	2011 en→fr	2012 fr→en	2012 en→fr	2014 fr→en	2014 en→fr	2015 fr→en	2015 en→fr
vanilla	24.6	23.4	28.8	26.6	30.7	28.7	31.3	30.9	31.1	29.5	36.0	36.0	36.2	34.0
BT	22.9	23.2	26.5	27.7	29.3	28.2	29.4	30.9	29.7	28.4	36.6	32.9	36.2	35.7
T-BT	24.5	23.8	28.9	27.3	31.2	30.0	31.8	31.6	31.8	28.9	37.3	38.2	36.6	36.0
selection	24.7	23.9	29.0	28.2	31.5	30.9	33.0	32.7	32.5	29.9	37.5	38.9	36.3	37.8

Table 6: BLEU scores for all the systems for en-fr on the overall test sets. "selection" denotes that decoding is performed by using the best model given the origin of the source sentence.