University of Zurich[UZH]

Bachelor's thesis
presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the degree of
Bachelor of Arts UZH

Probing Tasks for Noised Back-Translation

Author: Nicolas Spring
Student ID: 17-703-455

Examiner: Prof. Dr. Rico Sennrich

Supervisor: Mathias Müller MA

Department of Computational Linguistics

Submission Date: June 1, 2020

Abstract

When using back-translation, adding artificial noise to the synthetic source data leads to better model performance. This led to the hypothesis that noise can serve as an implicit label, signaling to the model that a given sentence has been back-translated. In this thesis, we use probing tasks, an adaptable model introspection technique, to verify this hypothesis, and we investigate the way in which the outputs of a model trained with explicitly tagged back-translation change when decoding a source sentence with and without an explicit label. We show that sentences with noise can be distinguished from genuine sentences with the information present in model states, thus confirming that noise can serve as an implicit label. Furthermore, we discover that decoding sentences with an explicit label yields translations that are lexically more diverse, while no clear changes can be observed in word order in respect to the source sentences. BLEU scores of hypotheses produced with an explicit label are lower than that of their standard-decoding counterparts, indicating that the interaction between lexical diversity and translation quality measured in BLEU is not yet fully understood.

Zusammenfassung

Bei der Anwendung von Back-Translation verbessert das Hinzufügen künstlichen Rauschens zu den synthetischen Daten die Qualität der Übersetzungen eines Systems. Dies führte zur Hypothese, dass der Effekt des Rauschens darin liegt, einem Übersetzungsmodell zu signalisieren, dass ein gegebener Satz künstlich generiert wurde. Um diese These zu überprüfen, benutzen wir Probing Tasks, eine anpassbare Methode zur Modellintrospektion, und wir untersuchen, inwiefern sich die Übersetzungen eines mit gekennzeichneter Back-Translation trainierten Modells verändern, wenn man das Decoding mit oder ohne diese Kennzeichnung durchführt. Wir zeigen, dass Sätze mit Rauschen von genuinen Sätzen anhand ihrer Repräsentationen im Encoder unterschieden werden können und bestätigen somit, dass das Rauschen als eine implizite Kennzeichnung fungiert. Zusätzlich stellen wir fest, dass das Decoding mit einer expliziten Kennzeichnung der Eingabesätze Übersetzungen von grösserer lexikalischer Diversität hervorbringt, aber kein eindeutiger Effekt auf die Parallelität der Übersetzungen zum Ausgangssatz festgestellt werden kann. Die BLEU Scores sind tiefer als die ihrer Gegenstücke mit normalem Decoding, was darauf hindeutet, dass die Wechselwirkung zwischen lexikalischer Diversität und in BLEU gemessener Übersetzungsqualität noch nicht in vollem Umfang erforscht ist.

Acknowledgments

I would like to thank Mathias Müller for his great enthusiasm, his constructive criticism, his many valuable suggestions and his very personal and cordial supervision during my work on this thesis. His impressive eye for detail and the essentials helped shape this thesis and he always lent a sympathetic ear in our constructive discussions.

I also want to thank Rico Sennrich for his enlightening insights and fruitful advice, which played a crucial role in shaping this thesis.

Many thanks as well to Arianna Bisazza, Myle Ott and Annette Rios for their rapid and thorough answers to my questions on their interesting work. The insights I gained were invaluable for this thesis.

Many thanks to Sean Murphey, Diego Gomes and Jason Weber for proofreading this text and their valuable feedback.

Finally, I want to thank my family for their support during this time. Their great understanding, enthusiasm, warm support and interest made this work possible.

Table of Contents

List of Figures

List of Tables

List of Acronyms

BLEU	Bilingual Evaluation Understudy
BT	Back-Translation
BPE	Byte Pair Encoding
CNN	Convolutional Neural Network
cMTLD	Copy-Aware Measure of Textual Lexical Diversity
cTTR	Copy-Aware Type/Token Ratio
EU	European Union
GPU	Graphics Processing Unit
LSTM	Long Short-Term Memory
MLP	Multilayer Perceptron
MT	Machine Translation
MTLD	Measure of Textual Lexical Diversity
NMT	Neural Machine Translation
RNN	Recurrent Neural Network
SMT	Statistical Machine Translation
TTR	Type/Token Ratio
WMT	Workshop on Machine Translation

1 Introduction

1.1 Motivation

With the advent of data-driven methods, the field of machine translation (MT) has changed significantly. Statistical machine translation (SMT) and neural machine translation (NMT) rely on large quantities of parallel data, from which they extract statistical knowledge. Neural machine translation models require training on millions of sentences to reach their full potential. However, depending on factors such as language pair, domain and cost, parallel data can often be difficult to obtain in large enough quantities. In almost all cases, monolingual data is more abundant. Thus, in recent years, an important element of research in the field of NMT has focused on making use of monolingual data to increase the performance of NMT systems. In contrast to SMT systems, there exists no single natural way to make use of this data in NMT systems, and many different approaches have been proposed in recent years.

One such approach is back-translation (BT; Sennrich et al., 2016a). In BT, monolingual data in the target language is translated to the source language using an intermediate translation model. This yields a parallel corpus with synthetic source sentences that can be used to train an NTM model. Intriguingly, adding artificial noise to this synthetic data (and thus unavoidably rendering some previously grammatical sentences ungrammatical) has been found to boost model performance even further (Edunov et al., 2018). It has been argued by Caswell et al. (2019) that this artificial noise serves as an implicit label and that explicit labels can perform this function more robustly. Since this phenomenon has previously been investigated only indirectly, by comparing the performance of different models, this research explores a more direct approach to determine whether genuine source text and backtranslations are indeed treated differently by an NMT model.

While neural networks can be effective in a variety of tasks, their inner workings lack transparency. Probing tasks offer a possibility of circumventing this difficult interpretability to gain insights into the encoding of the input. In a probing task,

model states are extracted and used as feature vectors for an auxiliary classification problem. If a classifier fails in predicting a label, it can be concluded that the model states do not store this information in a useful way (Adi et al., 2017). Therefore, probing tasks offer the possibility of directly inspecting how an NMT model treats BTs differently from genuine source text.

This thesis aims to answer the following research questions:

- Can genuine source sentences and noised back-translations be discerned with the information present in encoder states?
- How does explicitly labeling back-translations affect the output of an NMT model in terms of BLEU score, lexical diversity, and the distances between aligned words?

To that end, we train multiple MT models on parallel bitext and various kinds of synthetic source data. This includes decoding BTs by beam search, and adding artificial noise and explicit labels. To answer the first research question, we perform probing tasks with two models: A model trained on bitext and BTs obtained by beam search and a model trained on bitext and the same BTs as the first model, but with added synthetic noise. That way, we aim to answer the question of whether noised back-translated data can be recognized from model states. We test standard, beam-decoded, BTs to answer the question of whether BT is encoded differently in general or it is specifically noise that serves as the implicit label.

To shed more light on the practice of labeling BTs and its consequences for model outputs, we train an NMT model on explicitly labeled BTs and investigate how the output of the model changes with the presence of the tag in the input for genuine (and thus not back-translated) source sentences.

1.2 Outline

The remainder of this thesis is structured as follows: In Chapter 2, we review previous work relevant to BT and probing tasks. Chapter 3 discusses the data that was used for the training of NMT models. Chapter 4 describes our experiments and in Chapter 5, the results of our experiments are presented. We discuss the results in Chapter 6. Finally, Chapter 7 concludes this thesis. Appendix A lists the software that was used in the experiments and Appendix B includes the model hyperparameters.

2 Literature Review

In this chapter, we introduce literature relevant to our inquiry in order to contextualize our research and identify the gap in the research that we are aiming to fill. We review research on data augmentation with BT, model introspection and specifically probing tasks, as well as biases present in the output of machine translation models.

2.1 Back-Translation

Because they rely on statistical methods, the quality of statistical and neural machine translation models is heavily dependent on the available training data. This concerns both quality and quantity. In general, deep learning models perform better the more training data is available (Goodfellow et al., 2016, p. 19f). Therefore, large parallel corpora of sufficient quality are invaluable for the training of a translation model, but depending on factors such as language pair and domain, they are not always available. Typically, monolingual data exists in large amounts and for a variety of tasks (Lambert et al., 2011). In recent years, a lot of research has been dedicated to making this non-parallel data usable for training NMT models.

Back-translation (Sennrich et al., 2016a) allows for use of target-side monolingual data. In BT, the existing parallel training data is used to train a translation model in reverse direction (target language to source language). The target-side monolingual data is translated to the source language, which creates a parallel corpus with genuine target sentences and synthetic source sentences. This corpus is then combined with the parallel data to train a translation model from the source language to the target language, which means that the decoder is trained exclusively on genuine data, while the encoder is trained on a mix of genuine and synthetic data. Typically, beam search is used to obtain synthetic source sentences.

More recent research has aimed to better understand the way in which BT boosts the performance of NMT models. Imamura et al. (2018) have argued that with BT, the encoder and the attention do not benefit to the same degree as the decoder because their data is to a certain degree synthetic. Synthetic data may contain errors and

has in general less variability than human translations. For every target sentence, Imamura et al. generated multiple synthetic source sentences by random sampling, in contrast to the prevalent practice of using beam search. This boosted model performance considerably, which Imamura et al. attributed to increased source diversity. Edunov et al. (2018) have opted to generate only a single synthetic source sentence per target sentence. In addition to sampling, they use noising (Lample et al., 2018) to generate BTs. Their models trained on sampling and noising significantly outperform standard beam search. Edunov et al. have argued that this was due to a richer training signal. However, Caswell et al. (2019) suggested that the better performance of noised BT stems from signaling to the model that a given sentence was back-translated, thus serving as an implicit label. They generated BTs using the standard beam search approach and added a reserved `<BT>` token to serve as an explicit label. Model performance was on par with or better than the noising approach, indicating that a special tag can serve more robustly as an explicit label for BTs.

This thesis aims to further our understanding of how adding noise to back-translated data improves model performance. It explores whether NMT models treat noised back-translations differently from genuine source text and normal BT, utilizing probing tasks to determine whether noise serves as an implicit label.

2.2 Probing Tasks

The advent of neural machine translation has led to considerably improved quality, but also lower interpretability, of translation models. While statistical machine translation and rule-based machine translation are transparent to a certain degree, there is no straightforward way to interpret an NMT model. Answers to why a model makes certain choices are buried in large matrices of real-numbered values (Koehn, 2017, p. 90), and these matrices and vectors are difficult for humans to interpret. It is, however, crucial to better understand neural networks, as lack of knowledge about the sources of their performance and shortcomings limits our ability to design better architectures (Karpathy et al., 2015).

There are various ways of gaining insight into neural models. T-distributed Stochastic Neighbor Embedding (Maaten & Hinton, 2008) performs dimensionality reduction and is the most popular method to visualize high-dimensional vectors (Shi et al., 2016). Dimensionality reduction allows for the projection of vectors into a lower-dimensional space, where patterns become observable for humans. This can be applied to model states or word embeddings and allows for model introspection but

is limited in the specificity of the questions that can be answered. Other methods are specific to a certain architecture. Karpathy et al. (2015) trained character-level long short-term memory (LSTM; Hochreiter & Schmidhuber 1997) language models and identify LSTM cells sensitive to certain human-recognizable features such as quotes. Li et al. (2015) worked with the Stanford Sentiment Treebank dataset and visualized the embeddings that different architectures produce for selected sentences and phrases.

Probing tasks are a generic approach to investigate how certain features of an input sequence are encoded in an NMT model. They were first introduced by Shi et al. (2016) and Adi et al. (2017). In a probing task, only the encoder of an NMT model is needed. Hidden states are extracted from the encoder and are used as feature vectors for an auxiliary classification problem. This means that probing tasks are agnostic with respect to the model architecture, as long as the encoder produces a vector representation of the inputs (Conneau et al., 2018). This allows for a large range of applications. In the classification tasks itself, a classifier is trained to predict a label, which is a certain property of the input. If the classification fails, it can be concluded that this information is not stored in the vectors (and thus model states) in a useful way Adi et al. (2017). Crucially, probing tasks do not require the information to be interpretable to humans and allow for precise research questions.

Probing tasks have been used on a variety of different tasks at the word level (Belinkov et al., 2017; Bisazza & Tump, 2018), sentence level (Adi et al., 2017; Conneau et al., 2018; Raganato & Tiedemann, 2018) or both (Shi et al., 2016). Typical applications of probing tasks are part-of-speech tagging, syntactic information (e.g., sentence length, constituent structure or word order), morphology (voice, tense or gender), semantics or named entity recognition.

Motivated by the great diversity of possible applications of probing tasks that has been demonstrated by various researchers, we apply probing tasks to the detection of back-translated input. Because the type of noise that we apply to BTs (see Section 4.3) results in potentially shorter sentences with changed word order and not naturally occurring reserved tokens, previous work focusing on sentence length, word content and word order is especially relevant to us. Adi et al. (2017) found that an LSTM network was clearly capable of encoding both word content and word order, with the probing task clearly outperforming a random baseline. The encoding of sentence length was explored by Raganato & Tiedemann (2018), who found that in a transformer model (Vaswani et al., 2017), information about sentence length vanishes after the third layer. Most previous work has focused on analyzing encoder states of recurrent neural networks (RNNs) or convolutional neural networks

(CNNs). Thus, the work of Raganato & Tiedemann is also relevant to this thesis because of their choice in model architecture. To the best of our knowledge, probing tasks have not yet been applied to identify noised BT. We perform two experiments on a sentence level, training classifiers to discern standard and noised BT from genuine input text, respectively. We also test two different methods for obtaining fixed-sized sentence representations from the encoder states of a transformer model: 1) averaging the encoder states from the varying number of time steps and 2) concatenating the states and padding them with zeros to the length of the longest sequence.

2.3 Biases of Neural Machine Translation

Text that has been translated from a source language into a target language has a variety of features distinguishing it from naturally produced text. This dialect is referred to as *translationese*. Translationese text is typically simpler, more normalized, and explicit (Baker et al., 1993), as well as influenced by the source language (Toury, 2012). This is true not only for human translations, but for machine translation as well. Toral (2019) found that in post-edited machine translations, characteristics such as simplicity, normalization and interference from the source language are more distinct than in human translations, and he argued that these characteristics are already present in the MT output.

Toral (2019) investigated the lexical diversity of post-edited text by calculating both its type-token ratio and its lexical density, which is the ratio between the number of content words and the number of words in total. He found that both metrics are lower on MT outputs than on human translation. Similarly, Vanmassenhove et al. (2019) concluded that MT results in a general loss of lexical diversity. The authors calculated a variety of metrics to quantify lexical diversity for different model outputs. Notably, Vanmassenhove et al. did not use byte-pair encoding (BPE; Sennrich et al. 2016b), which they note may disadvantage neural models.

For the automatic evaluation of machine translation outputs, there are a variety of metrics, the most widespread of which is BLEU (Papineni et al., 2002). BLEU relies on reference translations to judge a translation hypothesis, and as such, BLEU is sensitive to the origin of the reference translation(s).[1] As our second research question is concerned with BT and the impact of explicitly labelling it at decoding, the

[1]It has been shown by Freitag et al. (2020) that the use of multiple reference translations does not increase the correlation with human judgment when compared to the use of a single reference translation.

work of Freitag et al. (2020) is especially relevant. The authors proposed additional references for the newstest2019 test set and demonstrated that the paraphrased references correlate well with human judgment and do not underestimate systems augmented with BT, which is often the case for references with translationese artifacts.

For the first time, the test data of WMT19 consists of (human-)translated data in both directions, resulting in a test set in which both the source and the target side contain 50% genuine source text and 50% translationese. This has made it possible to better assess the impact of references on BLEU scores of models with various data augmentation techniques like forward- and back-translation (see Edunov et al. 2019 and Bogoychev & Sennrich 2019). Additionally, Bogoychev & Sennrich (2019) explored domain effects in MT data, which can even occur in text from the same domain (such as news texts), when texts in different languages talk about different topics of specific importance to a country or region. They concluded that the original language of BT data can result in slight domain differences, which can be used to predict the provenance of sentences.

2.4 Hypotheses

Caswell et al. (2019) argued that noise in BTs can serve as an implicit label, signaling to the NMT model that a given sentence has been back-translated. Thus, we hypothesize that when using probing tasks to examine the encoding of genuine and back-translated sentences, it is possible to train classifiers to discern noised BTs from genuine text, but not normal, beam decoded, BTs.

Additionally, Caswell et al. (2019) argued that BT induces both a harmful and a helpful signal and found a small drop in translation quality measured in BLEU when decoding genuine source text "as-if-BT," that is, with an explicit label marking the sentence as back-translated. To further investigate the effects of decoding with and without an explicit label, we analyze the outputs that were generated that way according to three metrics:

1. BLEU score
2. Lexical diversity
3. Parallelism to the source text

We calculate BLEU both on WMT reference translations and on the additional references by Freitag et al. (2020). We hypothesize that similarly to the results

of Caswell et al. (2019), the BLEU scores when decoding as-if-BT are lower on the normal references. We anticipate higher BLEU scores on the paraphrases from Freitag et al. (2020), as we expect the model to produce more diverse output with an explicit label. We hypothesize that the NMT model learns to suppress more lexically diverse output when the input is not tagged. Finally, following Göckeritz (2020, to be published), we calculate a measure to quantify parallel word order between source and target sentences. This is done to quantify interference between the source sentence and the translation. As Caswell et al. (2019) argued that the explicit tag leads to stronger MT biases, we expect the text obtained by as-if-BT decoding to be more parallel to the source sentence than its standard counterpart.

3 Materials

As this thesis is concerned with the explanations of the effectiveness of noised back-translation, we followed the methods of Edunov et al. (2018) and Caswell et al. (2019) to reproduce transformer models comparable to theirs. The datasets in Chapter 3 and the model configurations in Chapter 4 reflect this fact. An overview of the data can be found in Table 1.

3.1 Parallel Data

The parallel data for model training consisted of the publicly available WMT'18 data for the English–German news translation task. All data except for the ParaCrawl corpus was used. The raw data consisted of 5,932,619 parallel sentences, which we normalized and tokenized. A joint source and target BPE (Sennrich et al., 2016b) with 32,000 merge operations was learned on the combined training data. It was then applied to the corpus to divide words into subwords. Sentences longer than 250 words and sentences exceeding a source/target ratio of 1.5 were filtered out. This resulted in a parallel corpus with 5,187,275 sentence pairs.

During the training of the four translation models, a validation set consisting of 52,396 sentence pairs was used. This validation set was split off from the parallel

Type		# Sentences	# Tokens		# Subwords	
		EN/DE	EN	DE	EN	DE
Parallel	Training	5,187,275	128,877,810	122,528,764	142,714,549	148,091,427
	Validation	52,396	1,302,826	1,238,616	1,441,063	1,495,425
Monolingual		24,274,464		445,898,200		604,040,213

Table 1: Corpus statistics of the monolingual and parallel data.

data. The sentences were normalized and tokenized, and the same BPE model was applied.

We used newstest2017 to evaluate the NMT models (see Section 4.2 for the calculation of BLEU scores), and it also served as the test set for the probing task. To analyze decoding with and without an explicit label, we additionally used newstest2014 and newstest2019. We also calculated BLEU scores on the additional reference translations and paraphrases from Freitag et al. (2020), which are publicly available on GitHub.[1]

For model training, the training and validation data were prepared by binarizing using fairseq (Ott et al., 2019),[2] and a joint dictionary with two additional tokens (`<BT>` and `BLANK`; see Section 4.3) reserved for tagging and noising BTs was created.

3.2 Monolingual Data

The monolingual German data for the BT experiments consisted of newscrawl2007-2017, which was distributed with WMT'18. It consisted of 260,772,552 sentences. A random subsample of size 25,000,000 was taken, and duplicates were removed. This resulted in a monolingual dataset of 24,274,464 sentences, which were then normalized and tokenized, and the BPE model that was learned on the parallel training data was applied.

The BT data was binarized with fairseq and the joint dictionary obtained from the parallel data was used.

[1] `https://github.com/google/wmt19-paraphrased-references`

[2] `https://github.com/pytorch/fairseq`, the exact fork that was used is available at `https://github.com/nicolasspring/fairseq-states`

4 Methods

Code and instructions on how to reproduce our experiments and results are available on GitHub.[1]

4.1 Model Architecture and Hyperparameters

For this thesis, a total of four transformer NMT models were trained (see Figures 4 and 5 in Appendix B for the exact training commands):

- A reverse (German-to-English) model for back-translating the monolingual target sentences.
- An English-to-German model trained on genuine parallel text and the synthetic BT data obtained with beam decoding (henceforth called *beamBT*).
- An English-to-German model trained on genuine parallel text and the synthetic BT data obtained with beam decoding and additional noise (henceforth called *noisedBT*).
- An English-to-German model trained on genuine parallel text and the synthetic BT data obtained with beam decoding and added tags (henceforth called *taggedBT*).

Model training was performed using fairseq (Ott et al., 2019) with 16-bit floating point operations. All models used the Transformer Big architecture (Vaswani et al., 2017). Training configurations were largely identical to those used by Edunov et al. (2018) and the publicly available BT examples.[2] The transformer model had six encoder and decoder layers and 16 attention heads. Word representations had a size of 1,024, and the feed forward layers had an inner dimension of 4,096. A dropout rate of 0.3 was used. The embeddings for the encoder, decoder, and output were shared. For label smoothing, a value of 0.1 was used.

[1] `https://github.com/nicolasspring/bt-probing-tasks`

[2] `https://github.com/pytorch/fairseq/tree/master/examples/backtranslation`

The models were optimized using the Adam optimizer (Kingma & Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ and $\epsilon = 10^{-8}$. The learning rate was varied over the course of the training following the equation of Vaswani et al. (2017):

$$lrate = d_{model}^{-0.5} \cdot min(step\_num^{-0.5}, step\_num \cdot warmup\_steps^{-1.5}) \tag{4.1}$$

There were 4,000 warmup steps, during which the learning rate was linearly increased. After this, it was decreased proportionally to the inverse square root of the step number. The configured learning rate was 0.0007.

The reverse (German-to-English) model was trained for 30,000 updates on the parallel data. The best checkpoint was determined on the validation loss.

All English-to-German models were trained on a combined dataset consisting of the parallel sentence pairs and monolingual data with a back-translated source side. For the noisedBT and taggedBT models, noise and tags were added (see Section 4.3). The ratio of bitext to synthetic data was adjusted by upsampling the bitext by a factor of 16. The models were trained for 100,000 updates and the last 10 checkpoints were averaged.

The models were trained on a single Tesla V100 GPU using a batch size of 3,584 tokens. Gradients were accumulated from forward and backward passes, and updates were performed every 128 batches. This was done to simulate the training conditions in Edunov et al. (2018) with synchronous updates on 128 GPUs and the same batch size of 3,584 per GPU. Thus, the effective batch size was 458,752 tokens. Model training took approximately 120 hours for the reverse model and 425 hours for the models trained on BTs.

4.2 Model Evaluation

The models were evaluated on newstest2017 using BLEU (Papineni et al., 2002) as the evaluation metric. For all models, we report two BLEU score values, namely tokenized BLEU score[3] and (detokenized) sacreBLEU[4] calculated with the reference implementation by Post (2018). This allows for comparability to both Edunov et al. (2018) (who reported tokenized BLEU) and Caswell et al. (2019) (who reported detokenized sacreBLEU).

[3]see `https://github.com/pytorch/fairseq/blob/master/examples/backtranslation/tokenized_bleu.sh`

[4]BLEU+case.mixed+lang.en-de+numrefs.1+smooth.exp+test.SET+tok.13a+version.1.4.9

Dataset	Sentence
German Monolingual (Target)	Wer bür@@ gt , wird ge@@ wür@@ gt .
beamBT Source	Those who are gu@@ ar@@ ant@@ ors are stran@@ gled .
noisedBT Source	who gu@@ BLANK ar@@ ors are stran@@ . gled
taggedBT Source	<BT>Those who are gu@@ ar@@ ant@@ ors are stran@@ gled .

Table 2: The three generation methods for BT applied to a German target sentence. Noise and tags are added to the beam search translation.

4.3 Generating Synthetic Source Sentences

Using the reverse model, the monolingual German data was back-translated to English. This was done using a beam size of five. This synthetic source data was combined with the monolingual target data and was used to train the beamBT model. To train the noisedBT model, noise was added to the beam search BTs. The noising approach followed Lample et al. (2018) and Edunov et al. (2018). Noise was added using the noisy-text toolkit by Valentin Macé.[5] There were three types of noise applied to the tokens: They were deleted with a probability of 0.1, then replaced with the filler token `BLANK` with a probability of 0.1, after which a random permutation was added such that no token was moved more than three positions away from its original position. The synthetic training data for the taggedBT model consisted of the beam search BTs with an explicit label, the `<BT>` token, which is inserted as the first token of all synthetic source sentences.

Table 2 provides an overview of the three different generation methods for synthetic source data using an example sentence from the corpus.

4.4 Probing Task

4.4.1 Data

The classifiers for the probing tasks were trained on (a subset of) the parallel NMT training data. Twenty-five thousand sentence pairs were subsampled from the training set. The English side served as the genuine source text for both experiments,

[5] `https://github.com/valentinmace/noisy-text/`

Dataset	genuine vs. beamBT		genuine vs. noisedBT	
	genuine	beamBT	genuine	noisedBT
Training	25,000	25,000	25,000	25,000
Test	3,004	3,004	3,004	3,004

Table 3: The number of sentences per class in the training and test sets for both probing task experiments.

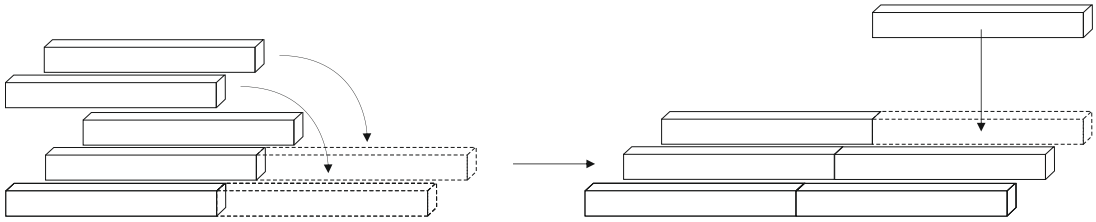
while the German side was back-translated with beam size five using the reverse model to yield the beamBT data. To obtain the noisedBT data, noisy-text was used to add synthetic noise to the sentences with the same values for deletion, replacement and permutation as in Section 4.3. This resulted in a balanced training set of 50,000 training examples for both experiments. It consisted of 25,000 distinct sentences each in two variations: 1) genuine source text and 2) a BT from the genuine target text (with added noise in the case of the noisedBT experiment).

The test set was prepared in a similar way to obtain a balanced test set with two variations for every input sentence. The English sentences from newstest2017 served as the genuine source text, while the German sentences were back-translated with beam size five using the reverse model. Noise was added using noisy-text and the same values for deletion, replacement and permutation as in Section 4.3. The test set contained 6,008 sentences: that is, the 3,004 sentences in newstest2017 in two variations (see Table 3 for an overview over the number of sentences per class in the datasets).

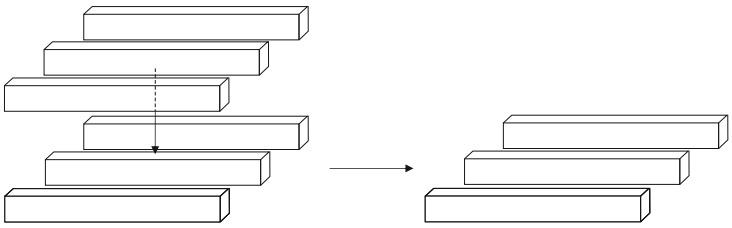
4.4.2 Experiments

We conducted experiments on two models. The genuine data and beam search BTs were encoded on the beamBT model to obtain data for the first probing task, whose aim was to discern genuine source text from beamBT. For the second probing task, whose aim was to discern genuine source text from noisedBT, the genuine data and the noised BTs were encoded on the noisedBT model. The encoder output was saved with pytorch.[6] All experiments were performed on the final output of the encoder; we did not extract any hidden states from previous layers.

[6] `https://pytorch.org/`



(a) Padding Method



(b) Averaging Method

Figure 1: Visualization of the averaging and padding methods to normalize the varying number of time steps per sentence. When padding, all sentence vectors are right-padded with zeros to the length of the longest sequence.

The 3D output matrices had dimensions for the time steps, the batch size and the encoder output size. They were subsequently converted to 2D feature matrices with one row per sentence. In our experiments, we tested two different representations for the source sentences, which differed in the way the varying number of time steps was normalized (see Figure 1 for a visualization of both methods). Our first method was concatenating the model states of the different time steps and right-padding them with zeros to the maximum number of time steps encountered in the training data. Henceforth, this method will be called *padding*. The maximum number of time steps in our experiments was 246; therefore, the feature vectors had a length of 251,904, that is, 246 times the encoder size of 1,024. The second method we explored was normalizing the sequence length by averaging the different model states (henceforth called *averaging*) for every input sentence. This resulted in feature vectors of length 1,024 and no padding was needed. For both methods, all batches were combined into single feature matrices containing the 50,000 or 6,008 sentences respectively.

For the probing tasks, both linear (sklearn `LogisticRegression` class) and non-linear (sklearn `MLPClassifier` class) classifiers were trained. The hyperparameters of the linear classifiers were identical to those of Bisazza & Tump (2018). For the non-linear classifiers, we used the Adam optimizer (Kingma & Ba, 2015). A random sample of 1,000 sentences of the training set was used as a validation set for early

stopping. The exact hyperparameters of our experiments can be found in Figures 6 and 7 in Appendix B. We report the accuracy on the test set extracted from newstest2017, as was described in Section 4.4.1.

4.5 Analyzing the taggedBT Model

We generated outputs with the taggedBT model both by decoding normally and by labeling the (genuine) source text as back-translated (henceforth called *as-if-BT*). As in training, this was done by adding the reserved `<BT>` token to the beginning of the source sequences. We report detokenized BLEU scores[7] with sacreBLEU (Post, 2018). For all test sets (newstest2014, -17 and -19), two translations were produced with the taggedBT model: One with unchanged source sentences and one with the added label. The BLEU was calculated with the WMT reference translation. For newstest2019, we additionally calculated BLEU scores with the references by Freitag et al. (2020). These include:

- AR: An additional reference translation
- WMT.p: A paraphrase of the original WMT reference
- AR.p: A paraphrase of the additional reference translation
- HQ(R): The best-rated reference translations for each sentence (taken from the WMT reference and AR)
- HQ(P): The best-rated paraphrases for each sentence (taken from WMT.p and AR.p)
- HQ(all): The best-rated references for each sentence (taken from the WMT reference, AR, WMT.p and AR.p)

To gain insights into the lexical diversity of the translations, we calculated two measures: Following the work of Toral (2019) and Vanmassenhove et al. (2019), we calculated the type-token-ratio (TTR) and the measure of textual lexical diversity (MTLD; McCarthy 2005) in their copy-aware versions (cTTR and cMTLD) that were proposed by Junczys-Dowmunt (2020) and used the Perl scripts provided by Junczys-Dowmunt on GitHub.[8]

Finally, to quantify the reordering of words in the translation, we trained a fast_align

[7] BLEU+case.mixed+lang.en-de+numrefs.1+smooth.exp+test.SET+tok.13a+version.1.4.9

[8] `https://github.com/emjotde/diversity`

(Dyer et al., 2013)[9] alignment model on the parallel training data, which we then applied to the test sets. We then extracted normalized distances between alignments similarly to Göckeritz (2020, to be published). We summed up the distance between the aligned words in a sentence pair. For example, if word 1 in sentence A was aligned with word 4 in sentence B, their distance was 3. We normalized the sum of distances by the number of alignments in a sentence. We then summed up the normalized distances for all sentences, which we normalized by the number of sentences. We performed this calculation on newstest2014, -17 and -19.

[9] `https://github.com/clab/fast_align`

5 Results

5.1 Model BLEU Scores

We report both tokenized and detokenized BLEU scores (see Section 4.2) for all models described in Section 4.1. Table 4 summarizes the model scores with tokenized BLEU and compares them to the BLEU scores reported by Edunov et al. (2018) for their setup with 24M back-translated sentences. Table 5 presents the detokenized BLEU scores of our models as well as the results reported by Caswell et al. (2019) for their setup with 24M back-translated sentences. Empty cells were not reported in the respective papers. Apart from the tokenized BLEU score of the beamBT model, all model scores were between 0.1 and 1.48 points lower than their respective counterparts in Edunov et al. (2018) and Caswell et al. (2019).

5.2 Probing Task

The results of the probing task are compiled in Table 6. Both experiments were successful in that both beamBT and noisedBT could be discerned by the classifiers better than chance-level or always predicting the majority class, which would have been 50%. For beamBT, the accuracy was between 56.18% (linear classifier with

Model	Language Pair	Tokenized BLEU	Tokenized BLEU (Edunov et al., 2018)
reverse	de-en	35.28	
beamBT	en-de	31.34	31.11
noisedBT	en-de	31.18	32.66
taggedBT	en-de	31.48	

Table 4: Tokenized BLEU Scores calculated on newstest2017. Scores reported by Edunov et al. (2018) with 24M back-translated sentences for reference.

Model	Language Pair	Detokenized BLEU	Detokenized BLEU (Caswell et al., 2019)
reverse	de-en	34.7	
beamBT	en-de	30.5	30.6
noisedBT	en-de	30.5	31.7
taggedBT	en-de	30.8	31.7

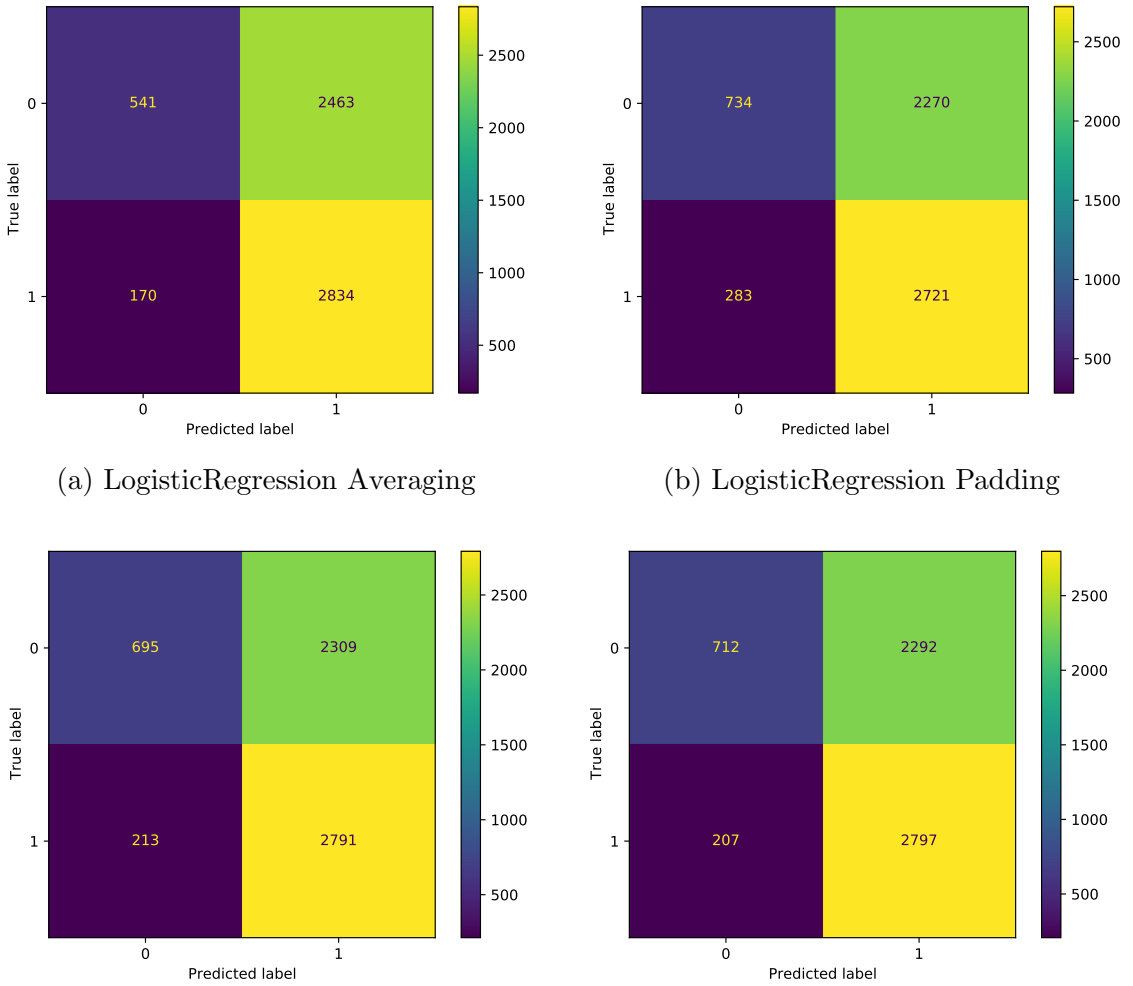
Table 5: Detokenized BLEU Scores calculated on newstest2017 using the sacreBLEU reference implementation. Scores reported by Caswell et al. (2019) with 24M back-translated sentences for reference.

Classifier	beamBT		noisedBT	
	padding	averaging	padding	averaging
LogisticRegression	57.51	56.18	98.14	**98.55**
MLPClassifier	**58.41**	58.02	98.02	98.42

Table 6: The accuracy of the classifiers in the probing tasks when predicting if an input is either genuine source text or back-translated (beamBT or noisedBT).

averaged states) and 58.41% (non-linear classifier with padded states). The accuracies for noisedBT were considerably higher, with the lowest accuracy of 98.02% being reached by the non-linear classifier on padded states and the highest accuracy of 98.55% being reached by the linear classifier with averaged states. For beamBT, the non-linear classifier performed a little better, while with noisedBT, the linear classifier reached the highest scores. Also, averaging performed a little worse on beamBT but better than padding on noisedBT. The very high results on noisedBT support our hypothesis. The lower, but still better than chance-level accuracies for beamBT do not allow us to directly rule out the possibility that (beam) BT is already discernible in and of itself.

Figures 2 and 3 present the confusion matrices for both experiments. Interestingly, in the beamBT experiments, all classifiers had a strong tendency to classify genuine source text as beamBT. In the noisedBT probing task, the effect was not as strong, and the overall accuracy was high.



(a) LogisticRegression Averaging

(b) LogisticRegression Padding

(c) MLPClassifier Averaging

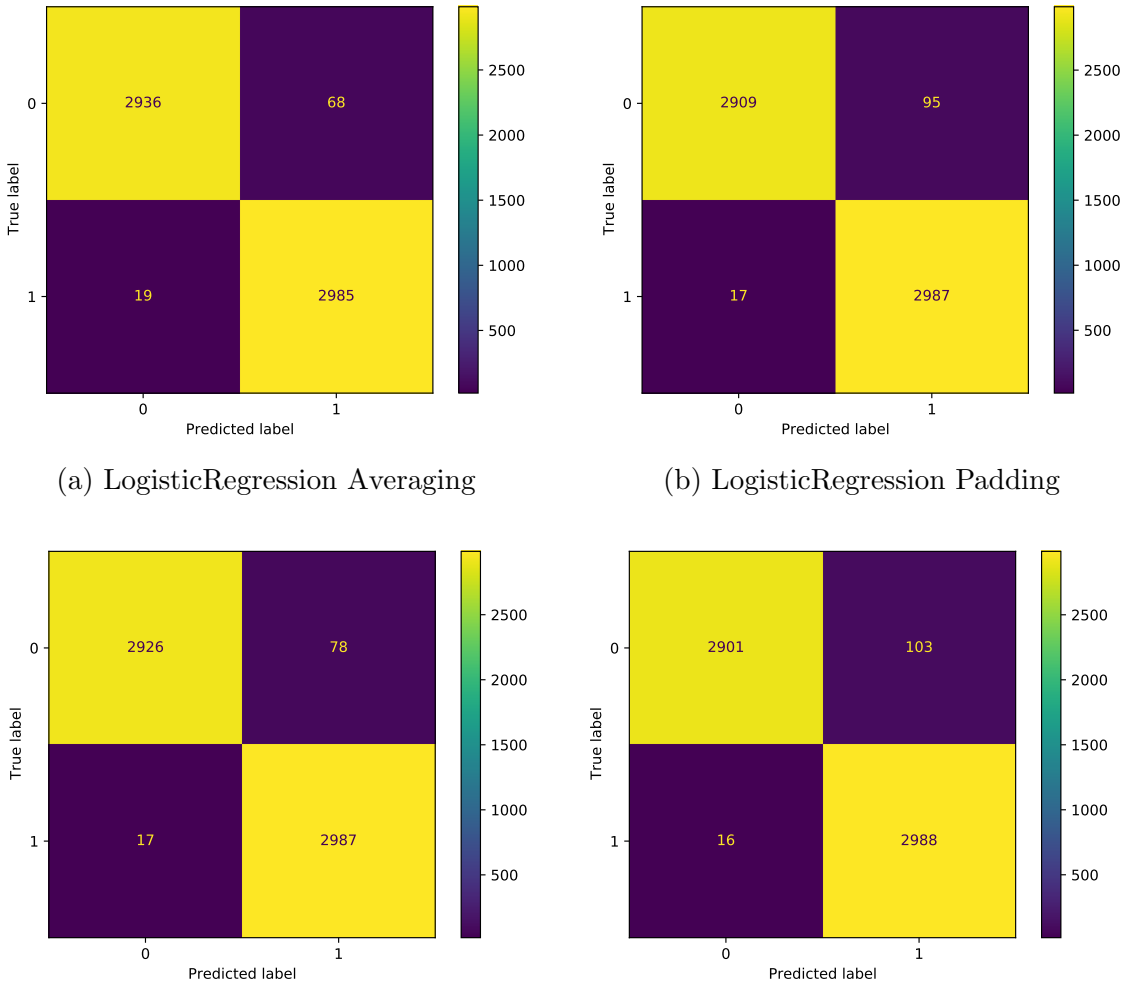
(d) MLPClassifier Padding

Figure 2: Confusion matrices of the beamBT probing task. Label 0 signifies genuine source text and label 1 signifies beamBT.

5.3 Decoding With and Without a Tag

5.3.1 BLEU Scores

The results of our BLEU evaluation of standard and as-if-BT decoding can be found in Table 7. In all our experiments calculating BLEU scores on model outputs, the as-if-BT hypothesis reached lower BLEU scores than the standard decoding hypothesis. This was true for the WMT references (newstest2014, -17 and -19) as well as the additional references by Freitag et al. (2020). While the lower BLEU scores on the normal references supported our hypothesis, the lower BLEU scores on the additional references were unexpected.



(a) LogisticRegression Averaging

(b) LogisticRegression Padding

(c) MLPClassifier Averaging

(d) MLPClassifier Padding

Figure 3: Confusion matrices of the noisedBT probing task. Label 0 signifies genuine source text and label 1 signifies noisedBT.

5.3.2 Lexical Diversity and Parallelism

Table 8 compiles the results for lexical diversity and parallelism. When exploring the values for cTTR and cMTLD, a clear pattern could be observed. For all test sets, the as-if-BT hypothesis achieved higher scores in both measures, meaning that the texts had higher lexical diversity. This supported our hypothesis. As the datasets differed in length (newstest14 2,737 lines; newstest2017 3,004 lines; newstest2019 1,997 lines), the values for these measures cannot be directly compared between different test sets, which is why we refrain from commenting on the differences between scores apart from the two decoding methods on the same dataset.

When comparing alignment distances, the picture is less clear. For newstest2014, the

Dataset	Standard Decode	as-if-BT Decode
newstest2014	**31.0**	30.4
newstest2017	**30.8**	29.8
newstest2019	**38.5**	34.4
newstest2019 AR	**31.7**	29.0
newstest2019 AR.p	**11.9**	11.3
newstest2019 WMT.p	**12.4**	11.7
newstest2019 HQ(R)	**35.0**	31.9
newstest2019 HQ(P)	**12.4**	11.7
newstest2019 HQ(all)	**27.3**	25.4

Table 7: Detokenized BLEU scores of the taggedBT model reached with standard and as-if-BT decoding on newstest2014, -17 and -19, as well as the additional newstest2019 references by Freitag et al. (2020).

Decoding Method	cTTR			cMTLD			Alignment Distance		
	2014	2017	2019	2014	2017	2019	2014	2017	2019
Standard	0.1961	0.1830	0.1791	93.8263	85.8923	74.4969	2.2136	1.9912	1.9615
as-if-BT	0.2023	0.1894	0.1848	97.4314	92.2911	80.0949	2.2014	2.0797	1.9775

Table 8: cTTR, cMTLD and normalized distance between aligned words calculated on newstest2014, -17 and -19.

score of standard decoding was higher, meaning that when aligning the hypothesis with the source, aligned word pairs were on average further apart than with as-if-BT decoding. For newstest2017 and newstest2019, this was not the case. Here, as-if-BT decoding achieved higher scores. The margin between the two values varied between the different test sets. It is important to note that the values for the alignment distances can be compared even with the varying length of the test sets, as the values were normalized by the number of words per line and the number of lines per document. Our results for alignment distances did not clearly support our hypothesis.

5.4 Conclusion

Both our probing tasks resulted in above chance-level accuracy. For beamBT, the accuracy was between 56.18% and 58.41%, and the classifiers reached an accuracy of between 98.02% and 98.55% for noisedBT. All beamBT classifiers showed a strong tendency to classify any input as beamBT. This effect was not as strong for the noisedBT class. We will discuss potential factors leading to this tendency in Chapter 6. When decoding with an explicit tag, our taggedBT model reached lower BLEU scores for all test sets, including calculation on the additional references by Freitag et al. (2020). In addition, we found that translations produced with an explicit label showed higher lexical diversity when measured in cTTR and cMTLD, but no clear effect could be observed for parallel word order in respect of the source sentence. In Chapter 6, we will additionally explore the connection between lexical diversity and naturalness, formulating possible explanations for this unexpected occurrence of both lower lexical diversity and lower BLEU score on the additional references by Freitag et al. (2020).

6 Discussion

6.1 Model BLEU Scores

As described in Chapter 5, the BLEU scores of our NMT models were lower than the scores of their respective counterparts in Edunov et al. (2018) and Caswell et al. (2019). We suspect that this is due to a non-optimal training setup in our part. A possible source of this disparity is the upsampling factor of the genuine bitext. Edunov et al. (2018) showed that higher upsampling values can adversely affect the output of NMT models. We did not conduct further experiments to identify the exact source of this disparity or optimize the models. All models had sufficiently high BLEU scores and were well trained to be comparable to the models described in Edunov et al. (2018) and Caswell et al. (2019). Thus, results from our models are likely also true for those respective models.

6.2 Noise as an Implicit Label for Back-Translations

Our experiments showed that beamBT and genuine bitext can be classified from model states with an accuracy of up to 58.41%. This was above chance-level and thus would suggest that NMT models do indeed treat this data differently. Comparing the accuracies to the noisedBT results (up to 98.55%; see Table 6), however, suggests that classifications could in theory be performed with a much higher accuracy. The setups for the experiments were comparable. We trained the same classifiers on a dataset which differed only on the BT part. The back-translated source sentences were obtained from a reverse translation, and for noisedBT, noise was added. The (forward) models that were used to extract the model states were also similar. They shared all hyperparameters, and the only difference at training was the back-translated input data. They also achieved comparable BLEU scores (see Tables 4 and 5). This comparability suggests that apart from the input data of the probing task itself, nothing should prevent much higher or indeed near-perfect accuracy (as was reached on noisedBT) in predicting beamBT from model states.

The fact that this is not the case indicates a disparity that must be explained and suggests that there is an inherent difference between the processing of beamBT and noisedBT data.

When examining the confusion matrices in Figure 2, we can observe a strong tendency to classify sentences as beamBT. This is in stark contrast to the confusion matrices of the noisedBT task (see Figure 3), where only a small number of genuine sentences were classified as noisedBT. We view this as evidence that beamBT in itself cannot be discerned reliably from genuine source text, as the above chance-level numbers for accuracy are the product of the nearly universally correct classification of beamBT and a much smaller number of true negatives in the classification of genuine source text.

We suspect that this bias when classifying the test set is rooted in other factors, specifically domain effects in our data. Before filtering, the raw training data for all NMT models consisted of the following corpora with varying sizes:

- Europarl v7: 1,920,209 lines
- Common Crawl Corpus: 2,399,123 lines
- News Commentary v13: 284,246 lines
- Rapid corpus of EU press releases: 1,329,041 lines

As can be seen from these numbers, the news texts only made up a small fraction of the parallel data. However, the monolingual data as well as the test sets consisted exclusively of news texts. It is thus a possibility that the models learned to pick up on these subtle domain differences and treat news texts differently from other text. Our training data for the probing tasks consisted of 25,000 sentences taken from the parallel data, which were back-translated with the reverse model and were added to the training set both in their original English form and as beamBT/noisedBT. Because the reverse model did not see much news data during training, we deem it possible that translating news data, which can be considered out of domain for this model, results in specific patterns. These patterns were carried over by the beamBT and noisedBT models when encoding, resulting in model states that are similar for both genuine and beamBT. Additionally, because the probing task classifiers were trained on the parallel data (only a small fraction of which was news texts), they were conceivably ill-equipped to perform the learned discrimination on the test data (exclusively news texts), which potentially contained different patterns. Notably, these patterns evoked by domain effects did not differ in both classes. This could explain the heavy skew towards a single class.

Intriguingly, in the noisedBT probing task, all classifiers still showed a slight tendency towards the noisedBT class, with the highest number of sentences being classified as genuine being 2,955 out of 6,008. We hypothesize that the same damaging signal may have also been present in this probing task. However, the two classes had more distinct features to begin with, thus allowing the classifiers to mostly ignore the damaging signal. This allowed for high accuracy on the probing tasks.

It can thus be concluded that our results are clear evidence that noise allows models to treat the input differently, which results in discernible model states. We also argue that despite our above-chance-level results on the beamBT task, beamBT cannot be discerned in and of itself. Thus, we argue that it is specifically the noise that allows for the classification of model states, the notion proposed by Caswell et al. (2019) that noise serves as an implicit label.

We also found differences in accuracy between the various combinations of classifiers and normalization methods for time steps. Further experiments would be needed to conclude whether these are general trends for this kind of probing task on a sentence level. As all configurations performed above chance-level and the differences in accuracy were small, it can be concluded that all configurations are valid to test for this specific research question.

6.3 Explicitly Tagging Back-Translations

6.3.1 Lexical Diversity

As shown in Table 8, the lexical diversity of hypotheses obtained by standard decoding was lower than that of hypotheses obtained by as-if-BT decoding. This is evidence for our expectation that the MT model can learn to suppress more diverse output when a label is not present. However, when calculating BLEU scores on the additional references by Freitag et al. (2020), the as-if-BT translations were rated lower, despite their higher values for cTTR and cMTLD. There are different ways of explaining this unexpected relation between higher lexical density and lower BLEU scores on the additional references, and until more experiments are performed, we can only speculate:

It is possible that decoding with a tag leads to higher lexical diversity, but not necessarily more natural output (note that Freitag et al. 2020 stated that specifically systems producing more natural output tend to be underestimated by references with translationese artifacts). The consequence of this would be that the concept

of lexical diversity is not necessarily indicative of naturalness. Also, while TTR and MTLD are well established metrics, they may not capture this phenomenon optimally, and there may be other measures which correlate better with BLEU on paraphrased references.

Another possibility is that the lexical diversity of the as-if-BT outputs was not pivotal for the lower BLEU scores. In most cases, the differences in BLEU between the decoding methods were small and anecdotal evidence from inspecting translations of a validation set suggests that both methods result in fluent and adequate translations. Our second hypothesis is thus that the lower BLEU scores cannot be entirely explained by the lexical diversity scores or may even be unrelated to them. This would suggest further experiments on phenomena not captured by lexical diversity measures, such as word order.

Regardless of the exact cause of the lower BLEU scores, decoding without a tag seems to adversely affect lexical diversity. Because high lexical diversity is regarded as a positive property, even separating genuine human text from translationese and human translationese from machine translationese (Junczys-Dowmunt 2020; also see Toral 2019 and Vanmassenhove et al. 2019), this can be regarded as an unwanted side effect of taggedBT.

6.3.2 Alignment Distances

Depending on the dataset, translations achieved both higher and lower scores when calculating the distance between aligned words. Our expectation that as-if-BT hypotheses would be more parallel was thus not clearly fulfilled. Our experiments suggest that there is no clear correlation between the decoding method and parallelism in respect to the source sentence when measured using the methods proposed by Göckeritz (2020, to be published). As of now, we cannot comment on the size of differences in word order our numbers indicate.

6.4 Limitations

From our experiments, we cannot conclude to what degree our probing task results were influenced by other factors, specifically domain effects. We deem it conceivable that domain effects were present in both the beamBT and the noisedBT experiments. Concerning the analysis of the taggedBT model, it is important to note that the phenomenon of lexical diversity cannot be equated with our two values for

cTTR and cMTLD. While TTR and MTLD are well-established, their copy-aware implementations have only recently been proposed by Junczys-Dowmunt (2020) and there is limited empirical knowledge on expectable values and the knowledge about the text that can be derived from such scores. Additionally, there exists a much wider range of metrics aiming to quantify the phenomenon of lexical diversity. Similarly, our calculation of alignment distances followed a method recently proposed by Göckeritz (2020, to be published), and there exists very little basic research on expectable values and the meaning of differing numbers for these distances. It is also important to note that the explanatory power of this metric is very dependent on the quality of the alignment model, which we did not empirically evaluate.

6.5 Implications and Future Work

While we can conclude that noise in BT allows a model to treat source sentences differently, our results suggest that the same is not true for beamBT. Further probing tasks identifying the domain of sentences would allow for a more decisive conclusion on the impact of domain effects on our beamBT results to be drawn. Bogoychev & Sennrich (2019) showed that similar domain effects could be used by an NMT model trained exclusively on synthetic data to identify whether a sentence stems from the source or target language "domain." Also, controlling for the source of both the training and test data would allow for our research question to be isolated and interfering factors controlled for, as it has been shown that even a random binary dummy feature added to half the data can lead to very high and thus overly optimistic results (Bisazza & Tump, 2018).

Identifying and further isolating the signal that allows an NMT model to detect noisedBT would lead to a better understanding of this phenomenon. As the analyzed noising approach consists of random permutations, deletions, and the replacing of tokens, conducting probing tasks with all these factors in isolation would allow for the determination of whether certain mutations are more impactful than others. Following the reasoning of Caswell et al. (2019), whose tags offer a less devastating possibility of labeling sentences as BT, isolating the most effective aspect of the noise could allow for the development of other noising approaches that could potentially be less devastating to the signal than "full" noise.

Based on our experiments, we suggest that further experiments on decoding with and without an explicit tag be conducted. This would allow for a better understanding of the impact that training with taggedBT has on the output of an NMT system. Additionally, experiments will have to be conducted to better quantify the alignment

distance values, allowing for the assessment of the practical meaning of the varying differences in score. We leave it to further research to conclude whether and to what degree the decoding method makes the hypothesis more parallel to the source sentence in respect to word order.

Our lower values for lexical diversity on standard decoding suggest that the advantage of BT systems is not fully exploited with taggedBT. At the same time, in our experiments, BLEU scores on the additional references by Freitag et al. (2020) were still higher with standard decode. There is a clear need for human evaluation and closer investigation of lexical diversity to better understand the source of this discrepancy. In addition, our results may contribute to a larger effort to determine the exact effects that the provenance of the reference can have on BLEU score. An optimistic goal of this research is to find techniques for the generation of references in a way that does not give an advantage to certain architectures or design choices (or controls for possible advantages in a different way).

7 Conclusion

Our first research question aimed to examine the notion proposed by Caswell et al. (2019) that noise in back-translated sentences can act as an implicit label signaling to the model that a given sentence has been back-translated. To achieve this, we used probing tasks, a generic tool for model introspection, to determine whether sentences can be correctly classified as genuine or back-translated from encoder states. We tested both standard (beam) BT and noised BT, using linear and non-linear classifiers and testing two techniques to normalize the number of time steps, averaging and concatenation. In our experiments, we were able to reach an accuracy of up to 98.55%, thus independently confirming the proposal of Caswell et al. (2019) using a different approach. We also conclude that normal (beam) BT cannot be discerned easily from model states and note a possible interference of domain effects, as our accuracy of up to 58.41% was reached by the classifiers having a strong tendency to classify sentences as beamBT regardless of their true label.

To answer our second research question, namely how explicitly labeling BT affects the output of NMT models, we conducted closer analysis of a model trained on tagged BTs. This technique has been deemed more effective than noised BT, producing state of the art results in Caswell et al. (2019). We aimed to better understand the effects that training with an explicit label for BT has on the output of both standard and as-if-BT decoding (decoding with an explicit label). This is because lower quality output for standard decoding could be potentially problematic if the model learns to hide positive aspects of BT when the explicit label is not present. We found some evidence for such behavior, as the standard decoding hypotheses reached lower values for both cTTR and cMTLD, two measures used to quantify lexical diversity. However, this did not result in lower BLEU scores. We used newstest2014, -17 and -19 as well as the additional newstest2019 references by Freitag et al. (2020), and on all references, standard decoding resulted in higher BLEU scores. Additionally, we found no clear evidence for as-if-BT decoding yielding references more parallel to the source sentence in terms of word order.

To conclude, we found strong evidence that noise in BT can serve as an explicit label. We deem it unlikely that the results were affected by normal (beam) BT al-

ready being discernible from genuine source text, as our classifiers reached accuracies much closer to chance-level on the test set and displayed a strong tendency towards classifying any input as beamBT. We suspect some interference from domain effects to be responsible for this lopsided classification. Concerning our first research question, our results confirm the notion proposed by Caswell et al. (2019) that noise in BT can serve as an implicit label. Regarding our second research question aiming at investigating the effects of training with tagged BT, we found evidence for lower lexical diversity when using standard decoding. This suggests that the practice of tagging BT has some unwanted side effects. However, the lower lexical diversity did not lead to lower BLEU scores on any of the tested references. In addition, no clear statement can be made if the decoding method affects how parallel translations are to the source in terms of word order. As BT itself can be regarded as a domain, we propose further experiments to investigate the interplay of domain and BT. We call for more research on the effects of lexical diversity on translation quality, both in terms of BLEU and of human-perceived adequacy and fluency.

References

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi & Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *Proceedings of ICLR Conference Track*, Toulon, France.

Baker, Mona, Gill Francis & Elena Tognini-Bonelli. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and technology: In honour of John Sinclair* 233. 250.

Belinkov, Yonatan, Nadir Durrani, Fahim Dalvi, Hassan Sajjad & James Glass. 2017. What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471* .

Bisazza, Arianna & Clara Tump. 2018. The Lazy Encoder: A Fine-Grained Analysis of the Role of Morphology in Neural Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2871–2876. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/D18-1313. `https://www.aclweb.org/anthology/D18-1313`.

Bogoychev, Nikolay & Rico Sennrich. 2019. Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation. *arXiv preprint arXiv:1911.03362* .

Caswell, Isaac, Ciprian Chelba & David Grangier. 2019. Tagged Back-Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, 53–63. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/W19-5206. `https://www.aclweb.org/anthology/W19-5206`.

Conneau, Alexis, German Kruszewski, Guillaume Lample, Loïc Barrault & Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

2126–2136. Melbourne, Australia: Association for Computational Linguistics. doi:10.18653/v1/P18-1198. https://www.aclweb.org/anthology/P18-1198.

Dyer, Chris, Victor Chahuneau & Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 644–648. Atlanta, Georgia: Association for Computational Linguistics. https://www.aclweb.org/anthology/N13-1073.

Edunov, Sergey, Myle Ott, Michael Auli & David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 489–500. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/D18-1045. https://www.aclweb.org/anthology/D18-1045.

Edunov, Sergey, Myle Ott, Marc'Aurelio Ranzato & Michael Auli. 2019. On The Evaluation of Machine Translation Systems Trained With Back-Translation. *arXiv preprint arXiv:1908.05204* .

Freitag, Markus, David Grangier & Isaac Caswell. 2020. BLEU might be Guilty but References are not Innocent. *ArXiv* abs/2004.06063.

Goodfellow, Ian, Yoshua Bengio & Aaron Courville. 2016. *Deep learning*. MIT press.

Göckeritz, Markus. 2020, to be published. *Eager Machine Translation*. University of Zurich MA thesis.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8). 1735–1780.

Imamura, Kenji, Atsushi Fujita & Eiichiro Sumita. 2018. Enhancement of Encoder and Attention Using Target Monolingual Corpora in Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, 55–63. Melbourne, Australia: Association for Computational Linguistics. doi:10.18653/v1/W18-2707. https://www.aclweb.org/anthology/W18-2707.

Junczys-Dowmunt, Marcin. 2020. Is MT really lexically less diverse than human translation? https://marian-nmt.github.io/2020/01/22/lexical-diversity.html. Retrieved May 14, 2020.

Karpathy, Andrej, Justin Johnson & Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078* .

Kingma, Diederik P. & Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980.

Koehn, Philipp. 2017. Neural machine translation. *arXiv preprint arXiv:1709.07809* .

Lambert, Patrik, Holger Schwenk, Christophe Servan & Sadaf Abdul-Rauf. 2011. Investigations on Translation Model Adaptation Using Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 284–293. Edinburgh, Scotland: Association for Computational Linguistics. https://www.aclweb.org/anthology/W11-2132.

Lample, Guillaume, Alexis Conneau, Ludovic Denoyer & Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043* .

Li, Jiwei, Xinlei Chen, Eduard Hovy & Dan Jurafsky. 2015. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066* .

Maaten, Laurens van der & Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(Nov). 2579–2605.

McCarthy, Philip M. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*: The University of Memphis dissertation.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier & Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, .

Papineni, Kishore, Salim Roukos, Todd Ward & 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*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. doi:10.3115/1073083.1073135. https://www.aclweb.org/anthology/P02-1040.

Post, Matt. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191. Belgium, Brussels: Association for Computational Linguistics. https://www.aclweb.org/anthology/W18-6319.

Raganato, Alessandro & Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 287–297.

Sennrich, Rico, Barry Haddow & 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)*, 86–96. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-1009. https://www.aclweb.org/anthology/P16-1009.

Sennrich, Rico, Barry Haddow & 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)*, 1715–1725. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/P16-1162. https://www.aclweb.org/anthology/P16-1162.

Shi, Xing, Inkit Padhi & Kevin Knight. 2016. Does String-Based Neural MT Learn Source Syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1526–1534. Austin, Texas: Association for Computational Linguistics. doi:10.18653/v1/D16-1159. https://www.aclweb.org/anthology/D16-1159.

Toral, Antonio. 2019. Post-editese: an Exacerbated Translationese. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, 273–281. Dublin, Ireland: European Association for Machine Translation. https://www.aclweb.org/anthology/W19-6627.

Toury, Gideon. 2012. *Descriptive translation studies and beyond: Revised edition*, vol. 100. John Benjamins Publishing.

Vanmassenhove, Eva, Dimitar Shterionov & Andy Way. 2019. Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, 222–232. Dublin, Ireland: European Association for Machine Translation. https://www.aclweb.org/anthology/W19-6622.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

CV

Personal Details

Nicolas Spring
Berglistrasse 4
8616 Riedikon
nicolas.spring@uzh.ch

Education

2017-2020	Bachelor of Arts in Computational Linguistics & Language Technology and Comparative Linguistics at the University of Zurich

Teaching Experience

2019	Teaching Assistant for the module *Maschinelle Übersetzung*

A Software

The code to reproduce the results is available on GitHub[1]. For data preparation and model training, the BT examples[2] on the fairseq repository were followed to a large part.

The following software was additionally used:

- A fork[3] of fairseq was used for model training and the saving of encoder states.
- noisy-text[4] by Valentin Macé was used to add noise to the data.
- diversity[5] by Marcin Junczys-Dowmunt was used to calculate the measures for lexical diversity.
- fast_align[6] was used to create word alignments between source and target texts.

[1] https://github.com/nicolasspring/bt-probing-tasks

[2] https://github.com/pytorch/fairseq/tree/master/examples/backtranslation

[3] https://github.com/nicolasspring/fairseq-states

[4] https://github.com/valentinmace/noisy-text/

[5] https://github.com/emjotde/diversity

[6] https://github.com/clab/fast_align

B Model Hyperparameters

Please refer to the following figures for the exact hyperparameters we used in our experiments. All hyperparameters not explicitly set by the training commands or hyperparameter dictionaries were set to the respective default values of fairseq and sklearn.

```
fairseq-train --fp16 \
    /path/to/binarized/data/directory \
    --source-lang de --target-lang en \
    --arch transformer_wmt_en_de_big --share-all-embeddings \
    --dropout 0.3 --weight-decay 0.0 \
    --criterion label_smoothed_cross_entropy --label-smoothing 0.1 \
    --optimizer adam --adam-betas '(0.9, 0.98)' --clip-norm 0.0 \
    --lr 0.001 --lr-scheduler inverse_sqrt --warmup-updates 4000 \
    --max-tokens 3584 --update-freq 128 \
    --max-update 30000 \
    --save-dir /path/to/checkpoint/directory
```

Figure 4: Hyperparameters of the reverse (German-to-English) model used to generate back-translations.

```
fairseq-train --fp16 \
    /path/to/binarized/data/directory \
    --upsample-primary 16 \
    --source-lang en --target-lang de \
    --arch transformer_wmt_en_de_big --share-all-embeddings \
    --dropout 0.3 --weight-decay 0.0 \
    --criterion label_smoothed_cross_entropy --label-smoothing 0.1 \
    --optimizer adam --adam-betas '(0.9, 0.98)' --clip-norm 0.0 \
    --lr 0.0007 --lr-scheduler inverse_sqrt --warmup-updates 4000 \
    --max-tokens 3584 --update-freq 128 \
    --max-update 100000 \
    --save-dir /path/to/checkpoint/directory
```

Figure 5: Hyperparameters of the three English-to-German models. Apart from the input data, the commands were identical.

```
from sklearn.linear_model import LogisticRegression

linear_clf_params = {'C': 0.0001,
                     'max_iter': 100,
                     'solver': 'liblinear',
                     'tol': 1e-4,
                     'verbose':100
                     }
linear_clf = LogisticRegression(**linear_clf_params)
```

Figure 6: Hyperparameters of the LogisticRegression model used for the probing tasks.

```
from sklearn.neural_network import MLPClassifier

mlp_clf_params = {'activation': 'relu',
                  'alpha': 0.0001,
                  'beta_1': 0.9,
                  'epsilon': 10e-8,
                  'hidden_layer_sizes': (100,),
                  'learning_rate_init': 0.0001,
                  'solver': 'adam',
                  'early_stopping':True,
                  'n_iter_no_change':10,
                  'validation_fraction':0.02,
                  'verbose':True
                  }
mlp_clf = MLPClassifier(**mlp_clf_params)
```

Figure 7: Hyperparameters of the MLPClassifier model used for the probing tasks.

Universität Zürich[UZH]

Philosophische Fakultät
Studiendekanat

Universität Zürich
Philosophische Fakultät
Studiendekanat
Rämistrasse 69
CH-8001 Zürich
www.phil.uzh.ch

Selbstständigkeitserklärung

Hiermit erkläre ich, dass die Qualifikations-Arbeit von mir selbst ohne unerlaubte Beihilfe verfasst worden ist und dass ich die Grundsätze wissenschaftlicher Redlichkeit eingehalten habe (vgl. dazu: http://www.uzh.ch/de/studies/teaching/plagiate.html).

Riedikon, 28.05.2020

.....

Ort und Datum Unterschrift