A Natural Diet: Towards Improving Naturalness of Machine Translation Output

Markus Freitag, David Vilar, David Grangier, Colin Cherry, George Foster
Google Research
{freitag, vilar, grangier, colincherry, fosterg}@google.com

Abstract

Machine translation (MT) evaluation often focuses on accuracy and fluency, without paying much attention to translation style. This means that, even when considered accurate and fluent, MT output can still sound less *natural* than high quality human translations or text originally written in the target language. Machine translation output notably exhibits lower lexical diversity, and employs constructs that mirror those in the source sentence. In this work we propose a method for training MT systems to achieve a more natural style, i.e. mirroring the style of text originally written in the target language. Our method tags parallel training data according to the naturalness of the target side by contrasting language models trained on natural and translated data. Tagging data allows us to put greater emphasis on target sentences originally written in the target language. Automatic metrics show that the resulting models achieve lexical richness on par with human translations, mimicking a style much closer to sentences originally written in the target language. Furthermore, we find that their output is preferred by human experts when compared to the baseline translations.

1 Introduction

Machine translation has made tremendous progress in recent years with the advent of neural methods (Bahdanau et al., 2015; Vaswani et al., 2017). This is especially true for language pairs with a large amount of available bilingual text for training (Barrault et al., 2020a). However MT output still can be improved: it currently trails human translators in expert evaluation (Toral et al., 2018; Freitag et al., 2021) and its language is perceived as poorer and more synthetic (Vanmassenhove et al., 2021). In this work, we aim to produce machine translation output that has a more *natural* style.

Source Es wird befürchtet, dass die Opferzahlen noch deutlich in die Höhe gehen.

Translationese It is feared that the number of victims will increase significantly.

Natural It is feared that the death toll will rise significantly.

Figure 1: Example De→En translations: This work sets the goal to generate more natural translations like *death toll/rise* in comparison to literal translations like *number of victims/increase*.

Although difficult to define precisely, we consider a translation to be natural if it is an adequate and fluent translation, whose style matches that of high quality monolingual text. Such a translation should contain few translationese constructs and use a rich vocabulary. This is exemplified in Figure 1. The translationese sentence uses the construct “number of victims”, which is a literal translation for the German “Opferzahlen”. Although correct (i.e. adequate and fluent), “death toll” shows a much more natural word choice for this translation.

Our objective in this paper is to study how the naturalness of machine translation output can be improved. In particular, we focus on how available measures can guide the translation process towards this goal. There have been several studies analyzing the naturalness of generated texts (see Section 2), but in contrast we concentrate on actively improving this aspect by modifying how NMT output is produced.

Our methodology follows a simple intuition: training data whose target side resembles high-quality text naturally written in the target language can bring model outputs closer to this style of text. We exploit the fact that bilingual training sets typically mix examples originating from both translation directions: source-to-target and target-to-source. We rely on contrasting language models (LMs) (Manning and Schütze, 1999; Moore and Lewis, 2010) to identify natural data: we train sep-

Findings of the Association for Computational Linguistics: ACL 2022, pages 3340 - 3353
May 22-27, 2022 ©2022 Association for Computational Linguistics

arate models on target-language data known to be translations, and on data known to be mostly originally written in the target language. We then use these LMs to tag parallel training data as having a natural or translated target side. Comparing to hard filtering of the data, tagging offers more flexibility without sacrificing coverage (Caswell et al., 2019).

Our contributions are as follows: (1) We use contrastive language model scoring to separate natural from translated text. (2) We demonstrate that optimizing BLEU scores on tgt-original test sets while avoiding high BLEU scores on src-original test set is a valid strategy to improve the naturalness of MT output. (3) We show that our more natural MT output is more similar to natural sentences based on lexical diversity. (4) Human evaluations show that the style of our more natural translations are preferred by humans, albeit with a minimal loss in translation accuracy.

2 Related Work

2.1 Translationese

Translations differ from text originally written in the target language due to a combination of factors that may include the intentional use of explicitation and normalization, or unintentional lexical or structural artifacts. The style resulting from the combination of these factors is often referred to as *translationese*. The effects of translationese in training data on MT quality and evaluation have been investigated by many authors (Kurokawa et al., 2009; Lembersky et al., 2012; Toral et al., 2018; Zhang and Toral, 2019; Graham et al., 2020; Freitag et al., 2019; Edunov et al., 2020; Freitag et al., 2020b). Several papers (Kurokawa et al., 2009; Koppel and Ordan, 2011; Shen et al., 2019; Riley et al., 2020) proposed to train classifiers to detect translationese sentences in monolingual corpora. Similar to our work, Kurokawa et al. (2009) used their classifier to preprocess MT training data, but they removed target-original pairs while we emphasize them. Lembersky et al. (2012) kept both types of data but introduced entropy-based measures that allowed their phrase-based decoder to favor lower entropy translationese entries. Riley et al. (2020) used a convolutional classifier to distinguish natural from translationese text. We train contrastive language models to partition the training data into original and translated sentences to bias the model to generate more natural translations.

2.2 Training Data Tagging for NMT

We use tags to differentiate subsets of the training data, with the objective of training a model that will decode differently depending on the tag provided at inference. This strategy has been explored with various objectives in prior work. Tagging to control inference has notably been used to indicate target language in multilingual models (Johnson et al., 2016), formality level (Yamagishi et al., 2016), politeness (Sennrich et al., 2016a), gender from a gender-neutral language (Kuczmarski and Johnson, 2018), backtranslation (Caswell et al., 2019), as well as to produce domain-targeted translation (Kobus et al., 2017). Shu et al. (2019) use tags at training and inference time to increase the syntactic diversity of their output while maintaining translation quality; similarly, Agrawal and Carpuat (2019) and Marchisio et al. (2019) use tags to control the reading level (simplicity/complexity) of the output.

2.3 Evaluation of Naturalness

Evaluation of MT usually focuses on accuracy and/or fluency (Barrault et al., 2020a; Läubli et al., 2020). Recent work has started to look at the richness and complexity of MT output. Vanmassenhove et al. (2019, 2021) address the effects of statistical bias on language generation. They assess lexical diversity and sophistication, and conclude that the translations produced by MT systems are consistently less diverse than the original training data, containing a higher number of frequent patterns while reducing the infrequent ones when compared to original texts. Toral (2019) compared MT output with human generated translations and found that there is a measurable difference between the two. In this work we use the diversity metrics introduced by Vanmassenhove et al. (2021) to demonstrate that we can build an MT system with lexical diversity similar to human translations (HT). We also incorporate the findings of Freitag et al. (2019), and show how to reliably evaluate more natural translations on target-original test sets while allowing the model to decrease BLEU scores on source-original test sets.

3 Approach

Our first objective is to distinguish text originally written in the target language (natural text) from translations. For that purpose, we train a pair of sentence-level language models to contrast their likelihood, a proven method for domain adaptation (Moore and Lewis, 2010; Axelrod et al., 2011). These language models are then used to tag MT training data as natural target (<nat>) or translationese target (<trans>), in order to train an MT system which can favor natural hypotheses.

3.1 Inferring Naturalness Tags

Our natural language model is trained on the monolingual newscrawl dataset from WMT (Barrault et al., 2020a). This data consists of web-crawled sentences from newspapers and other news sites from the countries speaking the corresponding language (e.g. Germany, Austria and Switzerland for German). Although it is not unusual to have contributions from foreign reporters or even translations of articles from foreign newspapers, we expect that the majority of the data collected this way will be natural text.

Our translationese LM is trained on machine-translated newscrawl data, as a proxy for human translated data. This approach does not require finding large amounts of existing text in the target language known to be translations, which is a challenging problem as the necessary metadata is not available for most corpora.

For our language models, we use a decoder-only transformer architecture comparable to *transformer-big* (Vaswani et al., 2017). We classify new sentences by thresholding the difference in average log probability under the two models.

For training our MT system we label each bilingual training example by prepending a special token in the source sentence denoting the class of the target sentence (<nat> or <trans>). At inference, we favor natural generation by prepending the natural token (<nat>) to the input. We call these models *natural-to-natural* (N2N) as their ultimate purpose is to translate natural source sentences into natural target sentences.

3.2 Potential Domain Biases

Domain bias might arise with our strategy. Our translated data originates from source language news and focuses on topics/domains of interest to a source-language speaker, while our natural data originates from target language news and therefore focuses on topics/domains of interest to a target-language speaker. When training a system that mainly concentrates on training data that originates from the target language, we might run into the problem that the model does extremely well on domains important in the target language while being poor on domains that are only important in the source language.

To counteract this problem, instead of relying on sentences originated in the target language only, we train on all the training data, but use tags to help the model learn the differences between the two training corpora. We then guide the inference algorithm (by using one of the two tags) to emphasize the characteristics important for one of the two training corpora only. The tagging approach helps the model to be familiar with the domains only important in the source language even when using a tag that emphasizes the characteristics of the target-original training data.

Finally, all human evaluations in this work are conducted with test sentences originating in the source language only, even when using the target-original tag. We will later show that humans prefer translations of sentences originated in the source language when using the target-original tag which demonstrates that by putting emphasis on the target-original training data, the model learns to translate better even though there is a mismatch between the domains of the training data and the test sentences.

4 Experimental Setup

We experiment on the WMT news translation tasks for evaluation (Bojar et al., 2016; Barrault et al., 2020b), focusing on the German↔English language pair. For this language pair there is abundant training data available, and MT systems achieve high quality translations. This is a good setting for our work since improving naturalness becomes a worthwhile endeavor only if high accuracy and fluency levels are reached.

4.1 Training Data

We use news-commentary-v15, paracrawl-v5.1, europarl-v10 and commoncrawl as training corpora (see Table 2). Noisy data is filtered out with contrastive data selection as proposed by Wang et al. (2018). Finally, we add back-translated data (Sennrich et al., 2016b) from the monolingual newscrawl (2007-2018) dataset for each target language, and mark synthetic source sentences with an additional special tag on the source side (<bt>) (Caswell et al., 2019). The BT data has been generated with a bitext only model from the reverse translation direction.

	Diversity Metrics											
model	B1↓	B2	B3↑	TTR↑	Yule's I↑	MTLD↑	H↑	D↓	PTF↓	CDU↓	SynTTR↑	cLM ↓
MT	68.55	6.31	25.13	0.1028	0.9375	144.07	12.64	92.22	0.7637	0.3938	0.1587	1.21
HT	68.25	6.30	25.44	0.1184	1.3980	148.08	12.93	92.00	0.7450	0.3781	0.1621	1.16
NAT	65.98	6.12	27.90	0.1553	2.9612	169.93	11.13	93.04	0.7133	0.3861	0.2108	0.77

Table 1: En→De: Diversity metrics calculated on the concatenation of newstest2011-2020 (∼25k sentences). HT scores are calculated on the src-orig half while NAT is calculated on the tgt-orig half. The cLM shows the ratio between the contrastive translationese and natural LMs. The arrows by the metric names indicate the desired behaviour towards more natural style. B2 does not have a clear desired behaviour.

4.2 Automatic Evaluation

4.2.1 Translation Quality

We use sacreBLEU (Post, 2018)[1] to automatically evaluate translation quality with BLEU, with the primary goal of improving scores on the *target-original* test sets. Since 2019, all WMT test sets have been composed only of source original (src-orig) sentence pairs. To create target original (tgt-orig) sets, we just flip the source and target of the test sets for the reverse direction. In previous years, the WMT test sets were a mixture of source- and target-original texts, each human-translated into the other language. For these years we split the test sets based on their original language and report results on the two subsets. Optimizing MT systems on these two settings can yield very different conclusions.

src-orig Beyond a certain level, BLEU scores on src-orig test sets are biased in favor of simpler and more literal translations (Freitag et al., 2020b); increasing scores above this threshold can have a negative impact on translation quality. Consequently, our goal is to avoid very high src-orig BLEU scores while increasing tgt-orig scores, a strategy that Freitag et al. (2020a) have demonstrated to be effective for improving translation quality.

tgt-orig Freitag et al. (2019); Edunov et al. (2020) found that MT systems trained with BT training data mostly improve on tgt-orig test sets. One explanation is that BT increases the fluency and naturalness of MT output, a property that can more easily be measured by comparing to natural target-language text than typical human translations, which have lower lexical diversity. Contrary to src-original test sets, generating literal, simple translation output decreases BLEU scores on tgt-orig test sets and cannot be used as a strategy to inflate BLEU scores. To further our main goal of generating more natural translations, we focus on improving BLEU scores on tgt-orig test sets.

4.2.2 Diversity Scores

Vanmassenhove et al. (2021) proposed a series of metrics to measure the lexical diversity of a text. They range from measures like type-to-token ratio (TTR) or the entropy of word forms given a lemma, to novel metrics that analyse synonym frequencies. They show that MT text has a lower degree of diversity than human-generated text but do not distinguish between original text and HT.

We refer the reader to the original paper for the metric definitions, although we also provide a short overview in the appendix. For better interpretability, in the results table we provide an indication of the desired direction for each metric. Note however that our goal is *not* to optimize these metrics, rather we want to build an MT system whose output is most similar to natural sentences. To illustrate this, assume we have a "translation model" that just generates random words. Such a system will certainly score high in diversity metrics (e.g. it will have a high entropy), but the resulting text will certainly not be natural. In fact, for a few metrics, our baseline system already gets a "better" score than natural sentences. Thus, for those metrics we should steer them in the "wrong" direction to achieve a style closer to natural sentences.

We used the implementation provided by the authors except for the "Synonym Frequency Analysis" metrics, which we reimplemented using an in-house synonym dictionary. Note also that some of these metrics are sensitive to the corpus size they are applied on (e.g. TTR, the type-to-token ratio, decreases as the corpus size increases). Thus not all numbers are in the same range as the results reported by Vanmassenhove et al. (2021).

[1] sacreBLEU signatures: BLEU+case.mixed+lang.LP +numrefs.1+smooth.exp+SET+tok.13a+version.1.5.1

4.3 Human Evaluation

We hired 4 professional translators (native in the target language) and conducted 2 types of human evaluations to evaluate (a) the overall translation quality, and (b) the naturalness of our MT output. We randomly chose 62 documents (roughly 1,000 sentences) from the src-original halves of newstest2019 for human evaluation to avoid human translated source sentences (Läubli et al., 2020).

Quality We measure quality with an in-context version of MQM (Lommel et al., 2014) which mimics the setup proposed by Freitag et al. (2021). This includes using the same error categories, severity levels and error weighting schema, which were adapted for the MT use case. As suggested in the study, we weight each major error with 5 and each minor error with 1, except for minor punctuation errors which get a score of 0.1.

Naturalness The preferred setup to evaluate naturalness is to present two translations of the same source sentence to native speakers without showing the source sentence. We ask the raters whether they prefer one of the outputs or rate them equally based on naturalness and natural phrasing. We emphasize that this evaluation is carried out in a monolingual manner, as showing the source can bias the human judges towards the translation that mimics the original sentence, as it is easier to evaluate.

4.4 Training Details

We train NMT models similar to the transformer-big (Vaswani et al., 2017) architecture (6 encoder and 6 decoder layers, model dimension size of 1,024, hidden dimension size of 8,192, 16 multi-attention heads). Our models use a vocabulary of 32k subword units (Kudo and Richardson, 2018) ands are trained for 250k updates with a batch size of 32k sentences. The baseline system uses only `<bt>` tags to tag all BT training examples while keeping the bitext data untagged. Our proposed system (denoted as N2N) is enhanced with the `<nat>` and `<trans>` tags to also tag the bitext data. During inference, in order to produce more natural output we tag the input sentence with the `<nat>` tag. For comparison purposes, we also analyze the output when using the `<trans>` tag.

5 Experimental Results

Due to space constraints and German being the more morphologically rich language, we focus our analysis mainly on the English→German (En→De) translation direction, but we provide translation results for the reverse direction (De→En) as well.

	size	NAT
news-commentary	251k	15.3%
commoncrawl	1.5M	39.6%
europarl	452k	44.1%
paracrawl	54.7M	30.4%
newscrawl-de	271M	92.0%*

Table 2: En→De: Training data statistics and fraction of natural target sentences. *This fraction is overestimated since this set is used for LM training.

5.1 Naturalness Classification

Our naturalness classifier contrasts the natural and translation LMs introduced in Section 3. We need to find a threshold to be able to classify the training data based on their target side as natural or translation. We chose 0.95 for both directions, resulting in $\sim$ 90% sentence-level classification accuracy on newstest2018. Table 1 (last column) shows the contrastive language model (cLM) scores for the concatenation of newstest 2011-20 for En→De for natural, (human) translated (HT) and machine translated (MT) sentences and shows that 0.95 seems a reasonable decision.

Table 2 reports the fraction of data classified as natural for each subset of the German side of our training corpus along with subset sizes. The fraction of natural target sentences per dataset varies between 30.4% and 44.1%, except for newscrawl-de (92.3%) which is our training set to define natural language and news commentary (15.3%) which mostly seems to have translations on the target side. The 44.1% of natural German sentences for Europarl is probably an overestimate and reflects the high quality of the translations in this particular corpus. Overall, the parallel corpora have less than 50% natural target sentences which means that the training data in this translation direction is dominated by translated text on the target side.

Table 3 shows the diversity metrics on a 15k sample of the training data. We can clearly see that the sentences considered natural are more lexically diverse than the sentences marked as translations, suggesting a valid classification by our model. Note that, as pointed out above, the lack of labelled data hinders reporting classification accuracy measures for the training data.

classified	B1↓	B2	B3↑	TTR↑	Yule's I↑	MTLD↑	H↑	D↓	PTF↓	CDU↓	SynTTR↑
TRANS	70.74	6.98	22.28	0.0918	0.9484	211.73	15.13	90.75	0.7296	0.3528	0.1328
NAT	68.91	7.16	23.93	0.1103	1.3501	303.05	14.94	90.69	0.7140	0.3726	0.1636

Table 3: En→De: Diversity metrics calculated on a 15k sample of the classified training data.

5.2 Translation Results

We evaluate three types of translations: the output of a regular baseline MT system and the outputs of our natural-to-natural (N2N) system trained with tags, decoding with either the `<nat>` or the `<trans>` tag. BLEU scores are reported in Table 4. We report average scores over all test sets (newstest 2011 through 2020), separate results for each set can be found in the appendix.

Focusing on En→De, for the src-orig half of the test sets, we obtain an average drop of 4.6 BLEU points when using the `<nat>` tag. For src-orig data, the references are translated text and the BLEU evaluation does not strongly reward text which does not adopt a translation style. When we instruct the system to produce translationese text using the `<trans>` tag, we recover the BLEU score of the baseline system. We thus have a clear indication that the system is learning to produce different texts depending on the given tag. This behaviour is consistent across all test sets, it is not just an effect due to averaging (see the Appendix for the detailed numbers).

We now turn our attention to the results on target-original data. In this situation the BLEU scores show a behaviour opposite to the previous case. Using the `<nat>` tag for translation, we get an improvement of 1.0 BLEU on average compared to the baseline. Remember that for this condition, the original text is on the target side, i.e. on the references we are evaluating against. This is thus an indication that we are indeed generating text that is closer to human natural text. When switching to `<trans>` translation, we see a drop of 2.4 points.

For the opposite direction we see a similar trend for both conditions (right part of Table 4).

5.3 Lexical Diversity Scores

In Section 5.2 we showed how BLEU scores change when applying our proposed method, and we observed an improvement on the target-original test sets, which may indicate improved naturalness in the output text. This evaluation setting is however artificial since it relies on translated source text while MT systems generally need to translate text originally written in the source language. We thus turn to a more detailed analysis of the produced translations, focusing on the src-original test sets.

Table 5 shows the diversity metrics computed on the concatenation of all the source-original test sets. It can be seen that the N2N system gets diversity scores much closer to ones calculated on natural sentences (NAT) when compared to the baseline system in all categories. In fact, it even obtains better scores than the human translations for some of them. We do not claim to outperform humans on translation quality: natural text shows certain characteristics that can be measured by these metrics, but improving on these metrics alone does not necessarily imply better translations. However, these results combined with the metrics from the previous section are positive indicators which motivate a human evaluation.

6 Human Evaluation

6.1 MQM

We carry out a human evaluation using the MQM framework (Lommel et al., 2014), which provides a detailed categorization of errors found in the text. The evaluation was carried out by professional translators. The results comparing the baseline output with the output of our N2N models with `<nat>` tag can be found in Tables 6 and 7.

Looking into the error categorization for En→De, we see a clear advantage of the N2N system for the style metrics, halving the number of major errors and reducing the number of minor errors by one third. The number of grammar errors has also been significantly reduced, from 56 minor errors in the baseline system to 29 in the N2N system, although with an increase of 6 major errors. For N2N we observe an increase in minor punctuation errors (mainly repetition of punctuation signs) and spelling errors, which can be traced back to the German orthography reform: the N2N seems to prefer the old writing form[2] which is now officially considered incorrect.[3]

[2] E.g. the N2N seems to generate more occurrences of "daß" instead of "dass".

[3] These errors could easily be corrected in a rule-based

		En→De		De→En	
		src-orig	tgt-orig	src-orig	tgt-orig
Base		38.0	37.0	36.4	45.4
N2N	`<nat>`	33.4	38.0	31.8	46.3
	`<trans>`	38.0	34.6	36.3	43.4

Table 4: Average BLEU scores for the WMT news datasets from 2011 to 2020.

Mode		B1↓	B2	B3↑	TTR↑	Yule's I↑	MTLD↑	H↑	D↓	PTF↓	CDU↓	SynTTR↑	cLM ↓
							En→De						
NAT		65.98	6.12	27.90	0.1553	2.9612	169.93	11.13	93.04	0.7133	0.3861	0.2108	0.77
HT		68.25	6.30	25.44	**0.1184**	**1.3980**	148.08	**12.93**	**92.00**	0.7450	**0.3781**	0.1621	1.16
Base		68.55	6.31	25.13	0.1028	0.9375	144.07	12.64	92.22	0.7637	0.3938	0.1587	1.21
N2N	`<nat>`	**67.48**	6.21	**26.31**	0.1099	1.1672	**156.19**	12.56	92.26	**0.7363**	0.3915	**0.1744**	**1.11**
	`<trans>`	68.53	6.32	25.16	0.1031	0.9446	145.88	12.72	92.17	0.7646	0.3948	0.1588	1.22
							De→En						
NAT		70.17	7.61	22.22	0.0835	0.7706	100.32	10.54	93.39	0.7888	0.3872	0.1847	0.83
HT		71.28	7.66	21.06	0.0878	0.6884	92.52	9.44	94.05	**0.7752**	**0.4194**	0.2431	1.14
Base		70.97	7.70	21.34	0.0982	0.8278	92.38	9.45	94.03	0.8023	0.4294	0.2399	1.25
N2N	`<nat>`	**69.88**	7.60	**22.53**	**0.1057**	**1.0220**	**98.49**	**9.76**	**93.84**	0.7813	0.4283	**0.2592**	1.14
	`<trans>`	70.97	7.71	21.32	0.0979	0.8235	93.42	9.46	94.02	0.8026	0.4280	0.2378	1.25

Table 5: En→De: Diversity metrics computed on the concatenation of newstest2011 to newstest2020, source-original test sets. Both the base and the N2N include backtranslated data. The arrows by the metric names indicate the desired behaviour towards more natural style. B2 does not have a clear desired behaviour.

For the accuracy errors, we also see an important reduction of mistranslation errors, from 79 to 26, but at the cost of increasing the number of major errors from 44 to 51. The other categories show comparable results between the two systems. Looking at the total number of errors, we see that the total number of errors decreases for the N2N system, from 508 for the baseline to 407 for the N2N system. The shift in errors is however not uniform across major and minor errors: while we achieve a drop of 30% in the number of minor errors (from 395 to 275), we increase the number of major errors by 16% (from 113 to 132). Overall, using the weighting approach proposed by (Freitag et al., 2021),[4] N2N achieves a better global score of 0.88, compared to 0.91 for the baseline system.

For the De→En translation direction, the results are mixed: we again obtain an important reduction in the number of minor style and grammar errors, but with with a slight increase of major errors. However the number of accuracy errors is also increased, which leads to a worse global score for the N2N system (0.49 vs. 0.55).

6.2 Side-by-side

The MQM analysis shows that the N2N system is able to produce grammatically better sentences, with some slight degradation in accuracy when compared with the baseline system. But, as pointed out before, a *natural* text might require more than grammatical and fluent text. In order to judge the naturalness, we carry out an additional evaluation where we present the translations produced by the baseline system and the N2N system to native speaker crowdworkers, and ask them to choose the better sounding one. Since MQM already judges the accuracy of the translations, this evaluation is *monolingual* and focuses solely on the naturalness of the sentences. Showing the source sentence may steer the human judges to choose the translation that is closer to it, as it is easier to judge, and we wanted to avoid this bias. The results can be found in Table 8. It can be seen that the human evaluators do have a preference for sentences generated by our N2N system. The difference is particularly important for the De→En translation direction. Some example translations are given in Table 9.

post-processing step.

[4] This weighting approach has been adapted for the machine translation use case, and differs from the standard weighting scheme used for human-produced translations.

	base		<nat>	
	M	m	M	m
Acc/Mistrans.	44	79	51	26
Acc/Omission	6	0	2	0
Acc/Addition	3	1	1	1
Acc/Untranslated	3	6	8	4
Fl/Grammar	14	56	20	29
Fl/Register	3	9	0	4
Fl/Inconsistency	0	2	1	0
Fl/Punctuation	0	57	2	72
Fl/Spelling	0	1	0	13
Fl/Display	1	10	8	4
St/Awkward	14	143	7	95
Ter/Inappr.	25	31	29	27
Other	0	0	1	2
Total Errors	113	395	132	275
Global Score	0.91		0.88	

Table 6: MQM scores for English-to-German of the baseline model compared to our N2N model with <nat> decode. The global score is a weighted combination of the error counts of all the categories. Lower scores are better. Major errors are under the 'M' column, minor errors under the 'm' column. Abbreviations are as follows: "Acc": Accuracy, "Fl": Fluency, "St": Style, "Ter": Terminology.

7 Conclusion

We propose a method for achieving more natural translations, i.e. translations which adopt a style closer to text originally written in the target language. Using contrastive language model scoring we classify our training data depending on whether the target side was originally written in the target language or whether it is a translation. This information is given to the translation system via an input tag, so that we can bias the generation process towards producing output closer to natural text. We demonstrate that building an MT system focusing on natural translations can be evaluated by optimizing BLEU on target-original test sets while avoiding high BLEU scores on src-original test sets. Through automatic metrics we show that the N2N method achieves lexical diversity closer to that of natural sentences indicative of more natural text. Indeed, human evaluations show that the produced translations are preferred by human judges when asked to choose the more natural translation. There is some drop in translation accuracy, as shown by the MQM analysis, however this can be an acceptable trade-off for some applications. For example, when considering post-editing, a more natural initial proposal will most certainly result in a more natural final output, while accuracy errors are usually easier to detect and fix for human post-editors.

The main contribution of this work lies in highlighting the potential for more natural translations by appropriate manipulation of the training data and evaluation measures. Our approach for using this information through tagging is a good first step, but it is a straightforward data manipulation. Other techniques that modify the model architecture or training objective may allow us to achieve the same improvements in naturalness without loss in translation accuracy.

	base		<nat>	
	M	m	M	m
Acc/Mistrans.	6	4	9	14
Acc/Omission	6	3	11	12
Acc/Addition	0	0	3	4
Acc/Untranslated	3	2	0	2
Fl/Grammar	1	31	4	9
Fl/Register	0	0	0	0
Fl/Inconsistency	4	3	2	2
Fl/Punctuation	1	3	5	4
Fl/Spelling	1	1	4	1
Fl/Display	0	0	0	2
St/Awkward	12	119	16	75
Ter/Inappr.	18	10	19	13
Other	0	0	0	0
Source Error	3	0	2	0
Locale/Date	0	1	0	0
Total Errors	55	177	75	138
Global Score	0.49		0.55	

Table 7: MQM scores for German-to-English of the baseline model compared to our N2N model with <nat> decode. Refer to Table 6 for a list of abbreviations.

Lang.	Preferences (%)			Num.
	<nat>	neutral	base	Ratings
EnDe	33.3	41.3	25.4	1000
DeEn	44.6	29.3	26.1	1000

Table 8: Human Evaluation: natural side-by-side of the baseline model compared to our N2N model with <nat> decode.

Source	Es wird befürchtet, dass die Opferzahlen noch deutlich in die Höhe gehen.
Baseline	It is feared that the **number of victims** will **increase** significantly.
N2N	It is feared that the **death toll** will **rise** significantly.
Source	Der Neubau sollte möglichst freundlich und hell gestaltet werden, damit sich die Bewohner darin wohlfühlen können, so der Architekt.
Baseline	The new building should be designed as friendly and bright as possible so that the residents can feel comfortable in it, according to the architect.
N2N	According to the architect, the new building should be made as friendly and bright as possible so that the residents can feel at ease in it.
Source	Musiker wie Janet Jackson, John Legend, Shawn Mendes und Cardi B haben bei einem gemeinsamen Konzert im New Yorker Central Park für mehr Engagement im Kampf gegen Armut und Krankheiten geworben.
Baseline	Musicians such as Janet Jackson, John Legend, Shawn Mendes and Cardi B have campaigned for more commitment in the fight against poverty and disease at a joint concert in New York's Central Park.
N2N	Musicians such as Janet Jackson, John Legend, Shawn Mendes and Cardi B joined forces at a concert in New York's Central Park to promote greater commitment to fighting poverty and disease.
Source	Bundesgesundheitsminister Jens Spahn hat sich für eine Neuregelung der Organspende ausgesprochen.
Baseline	Federal Health Minister Jens Spahn has spoken out in favour of a new regulation on organ donation.
N2N	Jens Spahn, Germany's Minister of Health, has called for a new regulation of organ donation.
Source	Grüß war schon vor zwei Jahren als damals 14-Jähriger in Bielefeld dabei.
Baseline	Grüß was already there two years ago as a 14-year-old in Bielefeld.
N2N	Grüß was in Bielefeld, Germany two years ago when he was 14 years old.

Table 9: Example translations for the German→English direction. The N2N translations have a more natural sentence structure when compared to the baseline translations. Further, N2N uses wordings that are more typically in natural written English text. For instance, when looking at the first examples: *number of victims* and *increase* are more literal translation than *death toll* and *rise* which are the more natural word choices in this context.

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564, Hong Kong, China. Association for Computational Linguistics.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020a. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors. 2020b. *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Ondrej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, and Lucia Specia. 2016. Ten years of wmt evaluation campaigns: Lessons learnt. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.

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, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, and Colin Cherry. 2020a. Human-paraphrased references improve neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1183–1192, Online. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, David Grangier, and Isaac Caswell. 2020b. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Vi'egas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *CoRR*, abs/1611.04558.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

James Kuczmarski and Melvin Johnson. 2018. Gender-aware natural language translation. *Technical Disclosure Commons*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, pages 81–88.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human-machine parity in language translation. *J. Artif. Intell. Res.*, 67:653–672.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265, Avignon, France. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM) : A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, pages 0455–463.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 193–203, Dublin, Ireland. European Association for Machine Translation.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. 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.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

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

Jiajun Shen, Peng-Jen Chen, Matt Le, Junxian He, Jiatao Gu, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. The source-target domain mismatch problem in machine translation. *arXiv preprint arXiv:1909.13151*.

Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. Generating diverse translations with sentence codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1823–1827, Florence, Italy. Association for Computational Linguistics.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

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 *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of* *the Third Conference on Machine Translation: Research Papers*, pages 133–143, Brussels, Belgium. Association for Computational Linguistics.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in Japanese-to-English neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 203–210.

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 Additional Results

A.1 Accuracy of Contrastive LM

The accuracy of the contrastive language model for all test sets for English→German are shown in Table 10. The accuracy is mostly around 90% for all test sets. In 2020, the test sets have been generated on the paragraph-level which could be the reason for the lower precision on the natural half. Some of the reference translations in earlier years have been post-edited from MT output which could be the reason why newstest2011 and newstest2013 have lower accuracy numbers for the natural sentences.

A.2 Per Test-set Results

Table 11 shows BLEU results for each separate test. It can be seen that all test set exhibit the same behaviour: increase tgt-orig and decrease in src-orig when using `<nat>`, the opposite for `<trans>`.

A.3 Results Without Backtranslation

Table 12 shows BLEU scores for the English→German translation direction, without using backtranslated data. We confirm that the N2N system using the `<nat>` also outperforms the baseline system on the tgt-original condition, while obtaining worse BLEU scores on the src-original evaluation. Using the `<trans>` tag, the score of the baseline system on the src-orig conditional is recovered.

Comparing the base system from Table 12 with the base system in the original paper, we see that the addition of backtranslated data, which is by construction natural on the target side, also behaves differently for the two evaluation conditions. Although it achieves improvements for both source and target original data, for the source-original condition it is only a minor improvement of 0.5 BLEU. On the other hand, for the target-original data we see a big gain of 3.1 points, further pointing towards the fact that the system generates more natural text.

B Short Overview of Diversity Metrics

In this section we provide a short overview of the diversity metrics used in this paper. For a full description, the reader is referred to (Vanmassenhove et al., 2021).

Lexical Frequency Profile (B1, B2, B3) The vocabulary is divided into three subsets: the 1000 most frequent words (B1), the next 1000 words (B2) and the rest (B3). The metric gives the percentage of running words in a text in each category.

TTR Type-to-token ratio, defined as the size of the vocabulary divided by the number of running words.

Yule's I Extension of TTR that is more robust to fluctuations due to text length.

MTLD Mean length of sequential words strings in the text that maintains a given TTR value.

H Shannon's entropy of word forms given a lemma.

D Simpson's diversity index of word forms given a lemma.

PTF Percentage of times the "primary" translation was chosen for source words with multiple translations.

CDU Cosine distance between the distribution of translation alternatives for a source word and a uniform distribution.

SynTTR Modified TTR limited to words with different translation alternatives.

	nt11	nt12	nt13	nt14	nt15	nt16	nt17	nt18	nt19	nt20	avg
NAT	77.9%	91.3%	67.8%	91.5%	91.8%	92.7%	88.1%	91.7%	93.0%	76.6%	86.2%
HT	81.4%	87.7%	79.4%	86.6%	85.7%	94.6%	87.2%	97.1%	93.6%	93.7%	88.7%

Table 10: English→German: Accuracy for all test sets.

		nt11	nt12	nt13	nt14	nt15	nt16	nt17	nt18	nt19	nt20	avg
Base		29.9	35.9	33.1	32.1	37.5	43.9	36.4	53.8	44.0	33.6	38.0
N2N	`<nat>`	27.4	31.5	30.9	30.0	34.0	36.1	32.0	44.5	37.8	29.9	33.4
	`<trans>`	30.0	35.2	33.3	32.4	37.9	44.1	36.3	53.1	44.1	33.1	38.0

(a) En→De: Source original side of test sets.

		nt11	nt12	nt13	nt14	nt15	nt16	nt17	nt18	nt19	nt20	avg
Base		33.3	33.5	42.8	37.5	31.7	39.4	32.8	45.8	41.8	31.1	37.0
N2N	`<nat>`	33.2	35.0	43.2	38.3	33.5	40.6	34.0	46.5	43.1	32.5	38.0
	`<trans>`	31.1	30.9	40.7	34.3	30.0	36.7	30.8	42.1	39.4	30.1	34.6

(b) En→De: Target original side of the test sets.

	mode	nt11	nt12	nt13	nt14	nt15	nt16	nt17	nt18	nt19	nt20	avg
Base		36.0	35.8	42.0	35.3	29.2	37.9	34.0	39.5	41.7	32.6	36.4
N2N	`<nat>`	33.5	32.7	36.9	29.9	25.3	32.5	30.3	33.9	34.5	28.3	31.8
	`<trans>`	36.3	35.2	42.1	34.9	29.2	37.8	33.6	39.8	41.7	32.7	36.3

(c) De→En: Source original side of test sets.

		nt11	nt12	nt13	nt14	nt15	nt16	nt17	nt18	nt19	nt20	avg
Base		39.2	44.1	39.6	39.9	44.0	54.9	47.2	58.6	48.2	38.1	45.4
N2N	`<nat>`	40.0	44.5	40.1	42.8	44.5	54.8	47.5	58.1	49.7	41.0	46.3
	`<trans>`	37.7	42.3	37.7	38.8	42.2	51.6	45.5	55.3	46.0	36.8	43.4

(d) De→En: Target original side of the test sets.

Table 11: BLEU scores for the WMT news datasets translation direction.

		nt11	nt12	nt13	nt14	nt15	nt16	nt17	nt18	nt19	nt20	avg
Base system		30.0	35.0	32.4	31.4	37.0	43.3	35.8	53.2	44.2	32.6	37.5
N2N	`<nat>`	28.0	31.1	29.9	29.2	33.3	36.4	31.7	44.9	38.1	29.7	33.2
	`<trans>`	29.9	34.9	32.5	30.9	36.8	43.2	36.1	53.7	44.4	32.0	37.4

(a) Source-original, no backtranslated data.

		nt11	nt12	nt13	nt14	nt15	nt16	nt17	nt18	nt19	nt20	avg
Base system		30.8	29.0	41.1	34.3	30.2	35.8	30.1	41.9	37.2	28.6	33.9
N2N	`<nat>`	32.2	31.3	42.6	35.7	30.9	36.6	31.0	42.6	38.6	29.6	35.1
	`<trans>`	30.5	29.2	40.3	32.0	28.3	34.0	28.4	39.2	36.1	27.9	32.6

(b) Target-original, no backtranslated data.

Table 12: BLEU scores for the English→German translation direction, without backtranslated data.