Pronoun-Targeted Fine-tuning for NMT with Hybrid Losses

Prathyusha Jwalapuram[*], **Shafiq Joty**[*§], **Youlin Shen**[*]
[*]Nanyang Technological University, Singapore
[§]Salesforce Research Asia, Singapore
[*]{jwal0001,srjoty,yshen010}@ntu.edu.sg

Abstract

Popular Neural Machine Translation model training uses strategies like backtranslation to improve BLEU scores, requiring large amounts of additional data and training. We introduce a class of conditional generative-discriminative hybrid losses that we use to fine-tune a trained machine translation model. Through a combination of targeted fine-tuning objectives and intuitive re-use of the training data the model has failed to adequately learn from, we improve the model performance of both a sentence-level and a contextual model without using any additional data. We target the improvement of pronoun translations through our fine-tuning and evaluate our models on a pronoun benchmark testset. Our sentence-level model shows a 0.5 BLEU improvement on both the WMT14 and the IWSLT13 De-En testsets, while our contextual model achieves the best results, improving from 31.81 to 32 BLEU on WMT14 De-En testset, and from 32.10 to 33.13 on the IWSLT13 De-En testset, with corresponding improvements in pronoun translation. We further show the generalizability of our method by reproducing the improvements on two additional language pairs, Fr-En and Cs-En.[1]

1 Introduction

The advent of neural machine translation (NMT) (Bahdanau et al., 2015; Vaswani et al., 2017) brought about significant improvements that left the previously successful statistical machine translation models far behind. However, the availability of large corpora has been no small part of that success, with recent NMT models using millions of sentences for training. A lack of availability of such large parallel corpora across languages has given rise to methods utilizing large amounts of monolingual data, such as for backtranslation (Sennrich et al., 2016a), language modeling (Çaglar Gülçehre et al., 2017; Zheng et al., 2020), or for large-scale pre-training (Lewis et al., 2020).

Backtranslation (Sennrich et al., 2016a; Edunov et al., 2018) is a commonly used strategy to improve MT models in the absence of adequate parallel data for training. A target-to-source model is first trained using the available parallel data, which is then used to translate a large target-monolingual corpus into the source to create pseudo-parallel data for training a source-to-target MT model. This has been shown to result in improvements in the BLEU score, and has become a popular method for improving NMT models, with many recent works proposing strategies to further improve it (Hoa; Yang et al., 2019; Caswell et al., 2019). However, recent studies have suggested that there is a limit beyond which the addition of synthetic data hurts the performance of the model (Fadaee and Monz, 2018; Poncelas et al., 2018). Also, recent work (Edunov et al., 2020; Nguyen et al., 2020) point out that back-translation suffers from the *translationese effect*, where back-translation only improves the performance when the source sentences are translationese but does not offer any improvement when the sentences are natural text.

Automatic post-editing (APE) is another common strategy that is used to improve translations. APE models are commonly monolingual, and typically take the output from some MT model as input, which they then modify. In the absence of adequate human post-edited data to train data-hungry neural models, Voita et al. (2019) and Freitag et al. (2019) both use round-trip translation data to train their post-editing models. In round-trip translations, target monolingual data is translated using a target-to-source model to the source text, and then back to the target using another source-to-target

[1] Code available at <https://github.com/ntunlp/pronoun-finetuning>.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2267–2279,
November 16–20, 2020. ©2020 Association for Computational Linguistics

model. This round-trip translated text is considered an approximation of poor quality MT output, which can be used in combination with the original target reference text to train the post-editing model. Voita et al. (2019) train a model to make corrections in context, using groups of 4 sentences as input, and show improvements in BLEU as well as translations of discourse phenomena.

NMT models typically fail on rare words that may not be adequately seen during training, such as named entities, or on words whose interpretation depends on the context such as discourse phenomena (Koehn and Knowles, 2017; Sennrich, 2018). For the latter, NMT models tend to prefer a more typical alternative to a relatively rare but correct one (*e.g.,* French "*Il*" is often wrongly translated to the more common "*it*" than "*he*"). However, these seemingly trivial errors can erode translation to the extent that they can be easily distinguishable from human-translated texts (Läubli et al., 2018).

There could be several reasons for why NMT models make such mistakes; our hypothesis is that since almost all NMT models are trained with a conditional language model objective, it is clear that this objective alone is proving inadequate to capture all of the information available in the text. We therefore propose a class of conditional generative-discriminative hybrid losses that explicitly teach models what to generate and what not to generate. Using these specialized losses, we aim to improve the learning power of the MT model.

Specifically, in this work, we target the improvement of pronoun translation by focusing our fine-tuning efforts through our proposed objectives and also through the fine-tuning data. We aim to leverage the training data we already have by extracting a subset of targeted fine-tuning data from the training corpus that the model has failed to learn correctly from. We use the newly proposed training objectives in combination with the targeted data to help the model fully reach its learning potential on the training corpus. We attempt to improve both general translation quality and the pronoun translation without compromising on either, and to do this without any elaborate model architecture.

Our main contributions are as follows:

- A class of Conditional Generative-Discriminative Hybrid losses that improve the learning potential of the model (§2).
- Effective fine-tuning strategy that uses the training data itself to improve MT (§3).
- Improvements in BLEU over WMT14 and IWSLT13 De-En testsets, and in pronoun translations over a pronoun challenge testset (§4.5, §5.3).
- Demonstration of generalizability through additional fine-tuning experiments on Fr-En and Cs-En (§5.4).

2 Targeted Finetuning Objectives

Before introducing our proposed Conditional Generative-Discriminative hybrid losses for fine-tuning NMT models on a targeted dataset, we first describe the Conditional Language Modeling (CLM) objective used to train NMT models.

2.1 Conditional Language Modeling

NMT models are generally trained with the CLM generative loss that relies on an auto-regressive factorization to perform density estimation and generation of target texts. For a source-target sentence pair (x, y), a CLM predicts a conditional probability distribution $P_\theta(y_{1:n}|x)$, where n is the number of tokens in the target text. The auto-regressive factorization for a CLM is given by

$$P_\theta(y_{1:n}|x) = \prod_{t=1}^{n} P_\theta(y_t|y_{<t}, \boldsymbol{c}) \qquad (1)$$

where $\boldsymbol{c}$ is a context vector that summarizes the relevant input (*e.g.,* attended vector over source text and the current decoder state). The CLM training objective for NMT can be written as:

$$\mathcal{L}_g = -\frac{1}{n}\sum_{t=1}^{n} \log P_\theta(y_t|y_{<t}, \boldsymbol{c}) \qquad (2)$$

Generating from CLM trained NMT models requires iteratively sampling from $P_\theta(y_t|y_{<t}, \boldsymbol{c})$, and then feeding y_t back into the model as input.

2.2 Generative-Discriminative Hybrid Loss

While CLM has been the de-facto loss to train NMT models, models trained with CLM make mistakes that can erode translation quality, making them easily distinguishable from human translation. For example, state-of-the-art NMT models are not very good at handling rare words like named entities. They have also been criticized for not being sensitive to discourse-level aspects such as pronouns, lexical consistency, and discourse connectives (Sennrich, 2018; Jwalapuram et al., 2020).

We introduce a generative-discriminative hybrid method for fine-tuning NMT models, with the motivation of generating tokens that are more strongly in one class vs. another. We consider that the reference tokens come from a positive class, whereas the model generated tokens come from a negative class. We propose two variants of our hybrid training – (*i*) log-likelihood and (*ii*) max-margin.

Log-likelihood training. Let $z \in \{0, 1\}$ represent the class for a training instance (x, y). We can consider a generative classifier as follows.

$$P_\theta(z = k|x, y) = \frac{P(z = k)P_\theta(y|x, z = k)}{\sum_{k'=0}^{1} P(z = k')P_\theta(y|x, z = k')} \quad (3)$$

Assuming an equal prior class probability, *i.e.,* $P(z = 1) = P(z = 0)$ and by replacing $P_\theta(y|x, z = k)$ with Equation (1), we can write:

$$P_\theta(z = k|x, y) = \frac{\prod_{t=1}^{n} P_\theta(y_t|y_{<t}, \boldsymbol{c}, k)}{\sum_{k'} \prod_{t=1}^{n} P_\theta(y_t|y_{<t}, \boldsymbol{c}, k')} \quad (4)$$

Since our objective is to maximize the probability of the reference tokens, we minimize the following negative log-likelihood loss:

$$\mathcal{L}_{nll} = -\log P_\theta(z = 1|x, y) \quad (5)$$

If y^+ is the reference (positive) translation and y^- is the model (negative) output, it is easy to show that the above loss is equivalent to

$$\mathcal{L}_{nll} = -\frac{1}{n}\sum_{t=1}^{n} \log \frac{\exp(\hat{y}_t^+/\tau)}{\left(\exp(\hat{y}_t^+/\tau) + \exp(\hat{y}_t^-/\tau)\right)} \quad (6)$$

where τ is the temperature parameter of the softmax,[2] and $\hat{y}_t^+$ and $\hat{y}_t^-$ are the final-layer logits (pre-softmax activations) corresponding to the reference token y_t^+ and model generated token y_t^-, respectively. The logit for the model generated token is computed by just taking the max over all the logits. We use $\tau = 0.5$ for our experiments.

Max-margin training. Following Collobert et al. (2011), we also propose a pairwise ranking loss that maximizes the distance between positive and negative samples. Formally,

$$\mathcal{L}_{mm} = \frac{1}{n}\sum_{t=1}^{n} \max\{0, \mu - \hat{y}_t^+ + \hat{y}_t^-\} \quad (7)$$

where μ is the margin; we use $\mu = 0.3$.

Note that the additional losses can be applied to all the tokens in the sequence, or restricted to some tokens. We demonstrate this in our experiments by applying the loss on all tokens and selectively applying the loss only on pronouns. Both of the discriminative losses essentially promote the probability of the positive (*i.e.,* correct) sample. However, the intuition behind using the additional loss over the standard loss is that the fine-tuning here focuses on improving the positive sample over the negative sample that the model has learnt to produce, rather than over the entire probability distribution over the full vocabulary.

We average these losses at both the sentence and the batch-level to add it to the existing CLM loss. The overall loss for training is

$$\mathcal{L}_{gd} = \lambda\mathcal{L}_g + (1 - \lambda)\mathcal{L}_d \quad (8)$$

where λ is a weighting hyperparameter, and the discriminative loss $\mathcal{L}_d$ is either $\mathcal{L}_{mm}$ (Eq. 7) or $\mathcal{L}_{nll}$ (Eq. 6). In our training, the discriminative loss $\mathcal{L}_d$ is aimed at correcting the mistakes, whereas the generative loss $\mathcal{L}_g$ is needed to preserve the translation adequacy and fluency. In our experiments, we simply set $\lambda = 0.5$.

3 Fine-tuning Data & MT Baselines

3.1 Pronoun-Targeted Fine-tuning Data

We create a subset of the training corpus in order to find training data that has not been fully learnt from; particularly, we focus our fine-tuning experiments on pronoun translation. Pronouns are an important discourse phenomenon that provide references to entities that have previously occurred in a text. Mistranslations can lead to loss of grammaticality or inference of the wrong antecedent, resulting in a misunderstanding of the text (Guillou, 2012).

Consider a parallel corpus $\mathcal{D} = (\mathcal{S}, \mathcal{R})$, where $\mathcal{S}$ is the source and $\mathcal{R}$ is the target/reference text. Assuming that the baseline NMT models (§3.2) are trained until convergence using this data, for our targeted fine-tuning of pronoun translations, we derive a subset of the training corpus $\mathcal{D}$ as follows:

1. Translate $\mathcal{D}$ using a baseline model $\mathcal{M}$ to obtain source to target translations $\mathcal{T}_\mathcal{M}$.
2. Align $\mathcal{T}_\mathcal{M}$ with reference $\mathcal{R}$ using *efmaral* (Östling and Tiedemann, 2016).
3. Find pronoun translations in $\mathcal{T}_\mathcal{M}$ that do not match reference $\mathcal{R}$. To exclude equivalent but non-identical translations, we use the list provided by Jwalapuram et al. (2019)[3].

[2] For the sake of simplicity, we omit τ in Eq. 3 - 4

[3] https://github.com/ntunlp/eval-anaphora

4. For each sentence with a mistranslated pronoun, extract the source sentences from $\mathcal{S}$.
5. The corresponding source and reference sentences form the pronoun-targeted fine-tuning subset, referred to as $\mathcal{D}_{\text{prn}} = (\mathcal{S}', \mathcal{T}')$.

3.2 Baseline MT Models

Typically, MT models are trained at the sentence level, taking one sentence as input and producing one sentence as output. Most MT systems at the sentence-level do not have access to adequate context that may be required for the translation of pronouns (Sennrich, 2018). Since it is our aim to improve pronoun translation, we train both a sentence-level model and a simple concatenation-based contextual model as our baselines:

SEN2SEN: A standard 6-layer base Transformer model (Vaswani et al., 2017) trained to translate each sentence independently.

CONCAT: A standard 6-layer base Transformer trained to translate a sentence given one previous sentence as context (Tiedemann and Scherrer, 2017). The input to the model is the previous sentence and the current sentence combined with a special separator character. Jwalapuram et al. (2020) show that this simple context model performs comparably or better than other elaborate contextual models like Voita et al. (2018), Zhang et al. (2018), and Miculicich et al. (2018).

Both the baseline models are trained for 100,000 steps. Other parameter details are in the Appendix.

4 Experiments

We conduct our fine-tuning experiments on the German-English (De-En) translation task. We describe our baseline training and fine-tuning corpus (§4.1), our experiments and results on fine-tuning using only the targeted subset data (§4.4), and fine-tuning using both the targeted subset data and the hybrid training losses (§4.5).

4.1 MT Training Data

Baseline training corpus. We use a De-En training dataset consisting of about 2.5 million sentence pairs, taken from the News Commentary, IWSLT (Cettolo et al., 2012) and Europarl (Tiedemann, 2012) corpora.[4] Sentences are encoded through Byte-Pair Encoding (BPE) (Sennrich et al., 2016b) with 40,000 operations, which results in a shared vocabulary of 40,224 tokens. We will refer to our baseline dataset as $\mathcal{D}$.

		WMT14	Pronoun Testset			
Model	**Train**	**BLEU**	**BLEU**	**P**	**R**	**F1**
SEN2SEN	$\mathcal{D}$	31.64	35.56	77.92	66.01	69.55
CONCAT	$\mathcal{D}$	**31.81**	**36.16**	**80.39**	**68.49**	**72.03**

Table 1: Baseline BLEU results on the WMT14 De-En testset and the BLEU (for translation), **P**recision, **R**ecall and **F1** scores (for pronoun translations) on the pronoun testset from Jwalapuram et al. (2019).

Pronoun targeted fine-tuning data. As described in §3.1, we derive the pronoun-targeted fine-tuning subset $\mathcal{D}_{\text{prn}}$ from the baseline training corpus $\mathcal{D}$ based on the translation errors of the baseline models. This results in a pronoun-targeted subset of 294,535 pairs for the SEN2SEN model and 285,783 pairs for the CONCAT model.

Random subset. We randomly extract a subset of 300,000 sentence pairs from $\mathcal{D}$, which approximately matches the size of the pronoun-targeted subset. We will refer to this dataset as $\mathcal{D}_{\text{rand}}$.

4.2 Pronoun Translation Evaluation

Testset. We run the models on the pronoun challenge testset provided by Jwalapuram et al. (2019), which is extracted from WMT testsets based on submission errors. For De-En, the testset has 2245 sentences, taken from WMT17-WMT19.

Evaluation. We report the macro-averaged F1 scores of the pronoun translation based on a simplified version of AutoPRF (Hardmeier and Federico, 2010). For each sentence in the testset, the counts of the pronouns in the system translation are clipped based on the pronouns in the reference translation; these counts are then used to compute the precision, recall and F1 scores.

4.3 Baseline Results

We first report the BLEU scores on the WMT14 De-En testset, and the BLEU, precision, recall, and F1 scores on the pronoun testset from Jwalapuram et al. (2019) in Table 1. The SEN2SEN model results in a BLEU of **31.64**, while the CONCAT model results in a slightly higher performance at **31.81** BLEU. More importantly, there is an improvement in the pronoun translations, with the F1 increasing from 69.55 for the SEN2SEN model to 72.03

[4] We exclude the UN corpus as our analysis showed that it does not have a high incidence of pronouns.

for the Concat model. To confirm that the context provides useful information rather than acting simply as a regularizer, we also run an experiment with the Concat model using a random sentence as context instead of the previous sentence. This model achieves a BLEU of 31.65 and a pronoun F1 of 69.65 - both lower than the baseline, confirming that the extended context from the previous sentence does provide helpful information.

4.4 Fine-tuning on Pronoun-Targeted Data

For the first set of fine-tuning experiments, we only fine-tune on the pronoun targeted subset $\mathcal{D}_{\text{prn}}$ for the Sen2Sen model. This helps us assess the training schedule so that we can achieve a balance between preserving the information from the full data and gaining targeted information from the subset.

Setup. Given a trained baseline model, we train additional epochs on the targeted subset $\mathcal{D}_{\text{prn}}$. Apart from training only on the subset data, we also try training on a shuffled dataset consisting of the training + targeted subset data (which essentially doubles the error-prone subset compared to the baseline training data), alternating the training between the subset and the full data ($\mathcal{D} + \mathcal{D}_{\text{prn}}$), and the subset and full data upsampled by 2 (*i.e.,* $2\mathcal{D} + \mathcal{D}_{\text{prn}}$).

To ensure that the results we see are from the fine-tuning and not simply from increased training, we train the original baseline model on the full data $\mathcal{D}$ for additional epochs, equivalent to the number of fine-tuning epochs.

Results. We see from the results in Table 2 that although the pronoun F1 improves, the BLEU scores drop when the model is fine-tuned only with the subset data $\mathcal{D}_{\text{prn}}$. Shuffling a mix of the full training data with the subset data leads to a smaller drop in BLEU and a gain in pronoun F1. However, alternating the training on the full corpus and the subset ($\mathcal{D} + \mathcal{D}_{\text{prn}}$) stabilizes the BLEU score, and upsampling the primary dataset ($2\mathcal{D} + \mathcal{D}_{\text{prn}}$) results in a smaller drop in BLEU, while gaining more significantly in pronoun F1 over the baseline. A similar trend is also observed for the Concat model. Further upsampling does not lead to a significant improvement in results, so all subsequent experiments upsample the primary dataset by 2.

Increased training of the baseline also results in a drop in BLEU scores. However, the pronoun F1 is also lower, which is not the case for the fine-tuning results, indicating that fine-tuning rather than increased training is the source of the improvements.

Fine-tuning	WMT14	Pronoun Testset			
data for Sen2Sen	BLEU	BLEU	P	R	F1
$\mathcal{D}$ (baseline)	31.64	35.56	77.92	66.01	69.55
$\mathcal{D}_{\text{prn}}$	30.43	34.72	79.49	67.55	71.02
$\mathcal{D} + \mathcal{D}_{\text{prn}}$ (shuffled)	31.31	35.48	78.35	67.02	70.35
$\mathcal{D} + \mathcal{D}_{\text{prn}}$	31.23	35.39	79.61	67.99	71.40
$2\mathcal{D} + \mathcal{D}_{\text{prn}}$	31.56	35.57	79.25	68.01	71.35
$\mathcal{D}$ (Increased training)	31.53	35.60	78.14	66.15	69.77
Concat					
$\mathcal{D}$ (baseline)	31.81	36.16	80.39	68.49	72.03
$2\mathcal{D} + \mathcal{D}_{\text{prn}}$	31.31	36.12	81.20	69.35	72.84

Table 2: Subset data: fine-tuning results on the WMT14 De-En with precision, recall and F1 scores on the pronoun testset. $\mathcal{D}$ represents the full training corpus; $2\mathcal{D}$ is the full training corpus upsampled by 2, while $\mathcal{D}_{\text{prn}}$ represents the pronoun targeted subset.

4.5 Effect of Additional Losses

We conduct experiments using both targeted data and proposed hybrid losses.

Setup. In accordance with our settings to alternate training between the upsampled full dataset and the subset data ($2\mathcal{D} + \mathcal{D}_{\text{prn}}$), we also alternate the additional loss such that it is only applied to the targeted subset. That is, in every alternate epoch, the model is trained on the upsampled full dataset ($2\mathcal{D}$) with the standard CLM translation loss $\mathcal{L}_g$ (Eq. 2), and then trained on the targeted subset $\mathcal{D}_{\text{prn}}$ with the proposed hybrid loss $\mathcal{L}_{gd}$ (Eq. 8).

Each fine-tuning model is trained for 9 additional epochs, such that the first and the last epoch use the targeted subset data and loss. This is effectively about 4 cycles of fine-tuning on ($2\mathcal{D}+\mathcal{D}_{\text{prn}}$); further training does not lead to improved loss.

Apart from applying the additional loss on all tokens in the subset data, we also experiment with applying the additional loss only on the pronoun tokens, *i.e.,* the loss is only applied to those tokens which have a pronoun as the target translation.

To further assess the contribution of the targeted subset data, we conduct experiments by replacing it with a random subset $\mathcal{D}_{\text{rand}}$. We also conduct fine-tuning experiments by applying the additional loss on the full training dataset $\mathcal{D}$ while training the baseline model for additional epochs.

Max-margin loss results. Results for fine-tuning with the max-margin loss are shown in Table 7a. We see that there is an improvement in BLEU from 31.64 to **32.14**. From the difference in improvement of the results from fine-tuning over

Model	Fine-tuning data	WMT14 BLEU	Pronoun Testset BLEU	 P	 R	 F1
Baseline SEN2SEN	-	31.64	35.56	77.92	66.01	69.55
Baseline CONCAT	-	31.81	36.16	80.39	68.49	72.03
All tokens						
SEN2SEN	$2\mathcal{D} + \mathcal{D}_{\text{prn}}$	**32.14***	36.16	78.83	66.15	69.77*
SEN2SEN	$2\mathcal{D} + \mathcal{D}_{\text{rand}}$	31.86	35.88	78.07	66.00	69.65
SEN2SEN	$\mathcal{D}$	31.75	36.34	78.27	66.36	69.91
CONCAT	$2\mathcal{D} + \mathcal{D}_{\text{prn}}$	31.75	**36.70**	**81.25**	**69.27**	**72.88**
Only Pronouns						
SEN2SEN	$2\mathcal{D} + \mathcal{D}_{\text{prn}}$	31.81*	36.43	78.62	66.82	70.37*
SEN2SEN	$2\mathcal{D} + \mathcal{D}_{\text{rand}}$	31.71	36.12	78.65	66.72	70.32
SEN2SEN	$\mathcal{D}$	31.89	36.20	78.31	66.32	69.98
CONCAT	$2\mathcal{D} + \mathcal{D}_{\text{prn}}$	**31.99***	**36.64**	**80.87**	**69.07**	**72.64**

(a) Fine-tuning results using **max-margin** loss.

Model	Fine-tuning data	WMT14 BLEU	Pronoun Testset BLEU	 P	 R	 F1
Baseline SEN2SEN	-	31.64	35.56	77.92	66.01	69.55
Baseline CONCAT	-	31.81	36.16	80.39	68.49	72.03
All tokens						
SEN2SEN	$2\mathcal{D} + \mathcal{D}_{\text{prn}}$	31.83*	36.50	79.18	67.16	70.78*
SEN2SEN	$2\mathcal{D} + \mathcal{D}_{\text{rand}}$	31.73	36.16	78.32	66.62	70.15
SEN2SEN	$\mathcal{D}$	31.77	36.24	78.35	66.17	69.86
CONCAT	$2\mathcal{D} + \mathcal{D}_{\text{prn}}$	**31.85**	**36.61**	**80.91**	**68.91**	**72.57**
Only Pronouns						
SEN2SEN	$2\mathcal{D} + \mathcal{D}_{\text{prn}}$	31.73	36.30	79.01	66.80	70.50*
SEN2SEN	$2\mathcal{D} + \mathcal{D}_{\text{rand}}$	**32.05**	36.43	78.35	66.25	69.87
SEN2SEN	$\mathcal{D}$	**32.05**	35.81	78.58	66.52	70.22
CONCAT	$2\mathcal{D} + \mathcal{D}_{\text{prn}}$	**32.00***	**36.57**	**80.89**	**68.66**	**72.39**

(b) Fine-tuning results using **log-likehihood** loss

Table 3: Targeted fine-tuning loss: fine-tuning results on the WMT14 De-En testset with F1 scores on the pronoun testset. Fine-tuning results on $2\mathcal{D} + \mathcal{D}_{\text{prn}}$ refer to alternated training with pronoun-targeted fine-tuning data and the upsampled full training data. Fine-tuning on $2\mathcal{D} + \mathcal{D}_{\text{rand}}$ is the same setting with the targeted data replaced with a random subset. Fine-tuning on $\mathcal{D}$ refers to additional training with the hybrid losses applied on the full dataset. * indicates statistically significant difference from the baseline ($p \leq 0.05$ for F1; $>80\%$ confidence for BLEU).

$\mathcal{D}_{\text{rand}}$ and $\mathcal{D}$, it is apparent that this increase is a consequence of both the targeted data and the targeted loss. There is also a corresponding increase in pronoun F1 from 69.55 to 69.77.

More importantly, we see that the CONCAT model drops slightly in BLEU to 31.75 with respect to the baseline, but the pronoun translation F1 improves from 72.03 to **72.88**. When the loss is applied only on pronouns, the SEN2SEN model has a smaller BLEU increase to 31.81, but a larger pronoun F1 increase to 70.37. The CONCAT model benefits the most from having both pronoun-targeted fine-tuning data and loss; both the BLEU score and the pronoun F1 improve.

Log-likelihood loss results. Results for fine-tuning with the log-likelihood loss are shown in Table 7b. The overall increase in BLEU with the log-likelihood loss is lower for SEN2SEN compared to the max-margin loss, but the improvements in pronoun F1 are higher. With respect to the results on fine-tuning over $\mathcal{D}_{\text{rand}}$ and $\mathcal{D}$, improvement in BLEU score here does not result in a corresponding improvement in pronoun translation, further confirming the contribution of the targeted data. Once again, the CONCAT model outperforms the rest by gaining in both BLEU and pronoun F1.

Both losses perform comparably - while the SEN2SEN model achieves a higher increase in BLEU with the max-margin loss, gains in pronoun translation are higher with the log-likelihood loss. For the CONCAT model, both losses provide similar BLEU improvements, but the max-margin loss leads to higher gains in pronoun F1.

5 Additional Experiments and Analysis

5.1 Qualitative Analysis of Results

We performed a qualitative analysis to see the effect of our fine-tuning. Some examples of improvements in translation resulting from our fine-tuning are shown in Table 4 (see Appendix for more).

The results of the targeted fine-tuning show that both the targeted data and the additional loss play a role in improving the translations. Another important conclusion that can be drawn is that there is no correlation between the BLEU score and the pronoun translation quality; in this case we have shown that it is possible to target the improvement of pronoun translations.

However, for the SEN2SEN model in particular, we see that there are improvements in BLEU that do not correspondingly improve pronoun translations, which can be surprising given that the fine-tuning data is targeted towards pronouns. It can be surmised from the improvements in the CONCAT model that the SEN2SEN model fails to improve in pronoun translation because it simply lacks the additional information that the context provides, which can be important for translating discourse phenomena like pronouns (Sennrich, 2018). See Table 4 for examples from the pronoun testset.

Another anomaly is that in some cases, the pronoun translation results are better when the loss is applied to all tokens rather than only to pronouns. A similar phenomenon may be the cause here - improved translation of the rest of the sentence may result in better contextual information, that in turn leads to better pronoun translations. This under-

Description	Examples
	WMT14 Testset
Source Reference Baseline Our best model	14 stunden kämpften die ärzte um das überleben des opfers , jedoch vergeblich . for 14 hours, doctors battled to save the life of the victim , ultimately in vain . 14 hours of doctors fought for the victim's survival , but in vain . the doctors fought 14 hours for the survival of the victim , but in vain .
Source Reference Baseline Our best model	der handel am nasdaq options market wurde am freitagnachmittag deutscher zeit unterbrochen . trading at the nasdaq options market was interrupted on friday afternoon , german time . trade at nasdaq options market was cut off on the german friday afternoon . trade in nasdaq options market was suspended on friday afternoon in germany .
	Pronoun Testset
Context Source Context Reference Baseline Our best model	... die die amerikanische flamme in die umnachtete welt bringe : lady liberty geht voran . sie soll die fackel der freiheit von den vereinigten staaten in den rest der welt tragen . ... taking the american flame out to the benighted world : **lady liberty** is stepping forward . she is meant to be carrying the torch of liberty from the united states to the rest of the world . it is meant to carry the torch of freedom from the united states to the rest of the world . she is supposed to carry the torch of freedom from the united states to the rest of the world .
Context Source Context Reference Baseline Our best model	versteinerte reste der haut bedecken noch immer die holprigen panzerplatten , die den schädel des tieres tragen . sein rechter vorderfuß liegt an seiner seite , seine fünf finger sind nach oben gespreizt . fossilized remnants of skin still cover the bumpy armor plates dotting the **animal's** skull . its right forefoot lies by its side , its five digits splayed upward . his right - hand front foot is on his side , his five fingers are spiked up . its right front foot is on its side , its five fingers are split upwards .

Table 4: Examples showing the improvements in translations from our best models, across the WMT14 and the pronoun testsets. The previous sentence context information for the pronoun testset is also shown.

	WMT14	Pronoun Testset			
Model	**BLEU**	**BLEU**	**P**	**R**	**F1**
Baseline Sen2Sen	31.64	35.56	77.92	66.01	69.55
DocRepair*	30.07	32.58	77.29	64.46	68.36
Backtranslation*	**32.57**	**38.54**	80.61	67.14	71.37
Best fine-tuned Sen2Sen	32.14	36.16	78.83	66.15	69.77
Best fine-tuned Concat	32.00	36.57	**80.89**	**68.66**	**72.39**

Table 5: Comparison with backtranslation and the DocRepair post-editing model. * indicates models use extra monolingual data. BLEU scores reported on the WMT14 De-En testset, with **P**recision/**R**ecall/**F1** on the pronoun testset. For DocRepair, the input is the output from our baseline Sen2Sen De-En model.

scores the importance of using context rather than trying to improve pronoun translations in isolation.

The general improvements in BLEU result from the fact that the targeted data is a subset that the model has failed to learn adequately from. Thus, our method of obtaining targeted data seemingly results in a subset that is generally poorly translated by the original baseline model, so training on this data results in an improved BLEU score. This also explains the disparity in results with the fine-tuning on the random ($\mathcal{D}_{\text{rand}}$) and the full ($\mathcal{D}$) datasets.

5.2 Comparison with Related Work

Backtranslation. We train a target-source En-De model with the same training data ($\mathcal{D}$, consisting of 2.5M pairs of parallel data) and settings as the baseline Sen2Sen model. This achieves a BLEU score of **27.4** on the WMT14 En-De testset. We use this model to translate about 76M sentences of NewsCrawl, a monolingual English corpus, to German. Using this pseudo-parallel corpus in addition to the original training corpus ($\approx$ 78M pairs), we train a Sen2Sen source-target De-En backtranslation model. This model is trained for 500K steps. The results are shown in Table 5. Although backtranslation achieves highest BLEU score at 32.57, our fine-tuned Concat model achieves the highest F1 for pronoun translation at 72.39, without having been trained on any extra monolingual data. This is further proof that it may be insufficient to simply improve the BLEU scores at a sentence-level. Performing fine-tuning on a Concat backtranslation model may be interesting to consider; we leave this for future work.[5]

Automatic post-editing. We train a contextual, monolingual automatic post-editing model proposed by Voita et al. (2019) for English. To capture MT errors, the model is trained with round-trip-translated texts as inputs with reference texts as the intended outputs. We use default settings and similar data sizes as proposed in their paper. We use 2.5M sentences from parallel data $\mathcal{D}$ and monolingual English sentences from NewsCrawl to make up $\approx$ 30M sentences. Using the En-De model described above and our baseline De-En model, we translate this data to German and then back to English to obtain round-trip translations. We use this

[5] A caveat here is that this would require training alternately on a targeted subset and an upsampled backtranslation dataset according to our training schedule. Considering the size of the backtranslation dataset, it would require massive amounts of additional training.

data to train their model[6] for around 750K steps as recommended by the authors.

We use the outputs of our baseline SEN2SEN De-En model on the WMT14 testset and the pronoun challenge testset as input to the model.[7] The results are shown in Table 5. We see that automatic post-editing does not lead to an improvement in BLEU[8] or pronoun translation in this case.

Our analysis of round-trip-translations suggests that this is possibly because they do not contain enough errors. Experiments conducted on the WMT14 En-De testset show that if it is translated using our En-De model (BLEU:27.40) to German and then translated using our De-En model (BLEU:31.64) back to English, the resulting text has a BLEU of 44.44, which is significantly higher. It is a well-known phenomenon that MT models perform substantially better on *translationese* (Graham et al., 2019), which refers to text that is unnatural by virtue of being translated. This means that it is not very likely to resemble typical MT output or capture the same errors (Poncelas et al., 2018); twice-translated texts therefore contain considerably fewer errors that can be learnt from.

	SEN2SEN		CONCAT	
Model	**BLEU**	**Prn. F1**	**BLEU**	**Prn. F1**
Baseline	31.64	60.47	32.10	62.01
Backtranslation	30.30	58.02	-	-
All tokens				
Max-margin	31.88	60.87	**32.95**	61.90
Log-likelihood	32.02	60.64	32.78	**62.10**
Only Pronouns				
Max-margin	32.13	60.61	**33.13**	**62.20**
Log-likelihood	32.16	60.83	32.78	61.97

Table 6: BLEU score and **Pron**oun translation **F1** results of the baselines and the fine-tuned models on the IWSLT13 De-En testset.

5.3 Results on the IWSLT13 Testset

We evaluate our fine-tuned models on the IWSLT13 De-En testset (Table 6). We also evaluate the pronoun translation for this testset. The backtranslation model fails to generalize, and performs worse than the baseline. It can be seen that our fine-tuned models improve over the baseline performance on this testset as well; the best SEN2SEN model improves from 31.64 to 32.16, while the best CONCAT model improves from 32.10 to 33.13, with corresponding improvements in pronoun F1. CONCAT continues to be the best performing model, showing significant improvements for both fine-tuning losses.

5.4 Generalizability to Other Languages

Finally, we test the generalizability of our fine-tuning method by running experiments for French-English and Czech-English. We use the same training dataset sources as for German-English (*i.e.,* News Commentary, IWSLT (Cettolo et al., 2012) and Europarl (Tiedemann, 2012)). This results in 2.53M sentences of training data and 500K sentences of fine-tuning data for Fr-En, and 992K sentences of training data and 100K sentences of fine-tuning data for Cs-En. We report the baseline BLEU results on the WMT14 testsets and the pronoun translation results on the corresponding testsets from Jwalapuram et al. (2019) containing 1478 (Fr-En) and 1686 (Cs-En) sentences. We see from Table 7 that our fine-tuning approach shows similar trends in improving BLEU and pronoun translation results for both Fr-En and Cs-En.

5.5 Discussion

Our objective is to propose a novel fine-tuning method that leverages "unlearned" data using additional loss. To this end, we proposed two different losses. We do not mean to advocate for any particular loss; in our experiments we happened to get comparable results, which may not conclusively point to one loss as being better. A different loss may perform better in other tasks.

Although we focused on pronoun translations, our fine-tuning method is generic and can be used to correct other kinds of errors in machine translations, like named entities or other rare words. Our proposed losses can be adapted to other directed generation tasks; *e.g.,* to improve coherence/factual correctness in abstractive summarization, or for controlled text generation. Our fine-tuning approach also opens up new ways to address training issues that originate from datasets; *e.g.,* it could potentially be used to correct biases (such as gender) or used to improve system robustness.

[6] Taken from https://github.com/lena-voita/good-translation-wrong-in-context.

[7] For the pronoun testset, we were only able to provide groups of 3 sentences as input instead of 4 which the original model uses, since the testset only provides two previous sentences as context. We add dummy text as the first sentence to make it a 4-sentence group input.

[8] Note that we calculate the BLEU scores for each sentence separately as is standard, unlike in groups of 4 as the original paper. This is to more accurately compare against the results from the rest of our experiments.

	Fine-tuning	WMT14	Pronoun Testset			
Model	loss	BLEU	BLEU	P	R	F1
Baseline SEN2SEN	-	35.61	34.53	90.64	64.00	73.73
Baseline CONCAT	-	36.06	35.18	84.86	72.07	75.86
All tokens						
SEN2SEN	max-margin	**36.12***	35.31	93.61	64.26	74.56*
SEN2SEN	log-likelihood	36.04*	35.39	96.39	66.95	**77.38***
CONCAT	max-margin	35.98	35.41	85.93	72.48	76.48
CONCAT	log-likelihood	35.98	35.09	85.07	71.43	75.51
Only Pronouns						
SEN2SEN	max-margin	36.05*	35.34	93.48	67.24	76.96
SEN2SEN	log-likelihood	35.86*	35.09	93.62	63.74	73.88
CONCAT	max-margin	35.97	35.26	85.71	71.97	76.07
CONCAT	log-likelihood	**36.09**	35.55	85.85	72.38	**76.50**

(a) Fine-tuning results for French-English

	Fine-tuning	WMT14	Pronoun Testset			
Model	loss	BLEU	BLEU	P	R	F1
Baseline SEN2SEN	-	25.23	21.88	82.65	48.78	60.40
Baseline CONCAT	-	28.27	24.19	71.94	55.57	60.37
All tokens						
SEN2SEN	max-margin	**26.13***	22.49	84.18	50.71	**62.16***
SEN2SEN	log-likehood	26.08*	22.65	83.02	49.02	60.53
CONCAT	max-margin	27.56	23.69	73.82	57.81	62.45*
CONCAT	log-likelihood	27.50	23.85	74.43	58.17	**62.89***
Only Pronouns						
SEN2SEN	max-margin	26.10*	22.56	83.02	49.96	61.03
SEN2SEN	log-likelihood	26.01*	22.62	83.90	49.17	60.88
CONCAT	max-margin	27.48	23.76	74.20	57.72	62.53*
CONCAT	log-likelihood	27.59	23.72	74.18	57.77	62.54

(b) Fine-tuning results for Czech-English

Table 7: Results for experiments on generalizability to other source languages, Fr-En and Cs-En. * indicates results are statistically significant.

6 Related Work

Our idea of conditional generative-discriminative training is related to the idea of discriminative training of generative models. Previously, this idea was proposed for Markov models. Collins (2002) trained a Hidden Markov Model (HMM) discriminatively for sequence tagging with structured perceptron algorithm. Yakhnenko et al. (2005) used a similar idea for sequence classification. In deep learning, the well-known generative adversarial networks (GANs) (Goodfellow et al., 2014) are an example where a generator is trained with the help of a discriminator. To the best of our knowledge, ours is the first work to explore this idea with conditional language models for guiding the model on what to generate and what not to generate.

A few fine-tuning methods are related to our work. Abdulmumin et al. (2019) pre-train an MT model on synthetic backtranslated data and fine-tune it on authentic parallel data, and show that it can improve 0.7 BLEU over backtranslation on English-Vietnamese. Fadaee and Monz (2018) use various sampling strategies to improve the results of backtranslation by targeting difficult-to-predict words based on prediction loss. Our strategy is similar in that we also try to target words that the model has trouble with, but we do not use additional data.

A number of methods have been proposed for adapting a trained MT model to another domain by fine-tuning. A common strategy is to simply perform additional training on the new domain dataset (Luong and Manning, 2015) or use a mix of in-domain and out-domain data for fine-tuning without loss of generalization (Chu et al., 2017) or upweight out-of-domain data (Wang et al., 2017).

There has been some work on targeted improvement of translations, specifically for named-entities. Ugawa et al. (2018) adapt MT network architecture to encode named entity features and tags while Li et al. (2018) perform domain adaptation in addition to feature encoding. With respect to discourse phenomena, Stojanovski and Fraser (2019) propose a curriculum learning based approach, where a context-aware model is trained on randomly sampled oracle data containing gold-standard pronouns. In our work, we focus on the baseline model's failings and try to increase its learning capacity by proposing additional losses.

Most recent work on improving pronoun translations has involved building more complex architectures that incorporate contextual information (Voita et al., 2018; Wong et al., 2020). In contrast, we present a more generalized approach.

7 Conclusions and Future Work

We have proposed a class of conditional generative-discriminative losses to increase the learning potential of NMT models, showing that it is possible to leverage "unlearned" training data to further improve an MT model, by strategically filtering the data and applying additional targeted losses.

We demonstrated the effectiveness of our methods on different languages and testsets, also reporting improved pronoun translations. Although we focus on pronoun translations, our fine-tuning method is generic and can be used to correct other kinds of errors in machine translations, like named entities or other rare words. In future work, we will explore other such applications of our proposed methods.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. Shafiq Joty would like to thank the funding support from NRF (NRF2016IDM-TRANS001-062), Singapore.

References

Idris Abdulmumin, B. S. Galadanci, and Aliyu Dadan Garba. 2019. Tag-less back-translation. *ArXiv*, abs/1912.10514.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ArXiv*, abs/1409.0473.

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.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of simple domain adaptation methods for neural machine translation. *ArXiv*, abs/1701.03214.

Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 1–8. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

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

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 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.

Marzieh Fadaee and Christof Monz. 2018. Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. Text repair model for neural machine translation. *ArXiv*, abs/1904.04790.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *ArXiv*, abs/1906.09833.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10, Avignon, France. Association for Computational Linguistics.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Comput. Speech Lang.*, 45:137–148.

Christian Hardmeier and Marcello Federico. 2010. Modelling pronominal anaphora in statistical machine translation. In *Proceedings of the 2010 International Workshop on Spoken Language Translation*, IWSLT '10, pages 283–289, Paris, France.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. 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 2964–2975, Hong Kong, China. Association for Computational Linguistics.

Prathyusha Jwalapuram, Barbara Rychalska, Shafiq R. Joty, and Dominika Basaj. 2020. Can your context-aware MT system pass the DiP benchmark tests? : Evaluation benchmarks for discourse phenomena in machine translation. *ArXiv*, abs/2004.14607.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zhongwei Li, Xuancong Wang, Ai Ti Aw, Eng Siong Chng, and Haizhou Li. 2018. Named-entity tagging and domain adaptation for better customized translation. In *Proceedings of the Seventh Named Entities Workshop*, pages 41–46, Melbourne, Australia. Association for Computational Linguistics.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. In *Advances in Neural Information Processing Systems 34*. Curran Associates, Inc.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Alberto Poncelas, Dimitar Sht. Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. *ArXiv*, abs/1804.06189.

Rico Sennrich. 2018. Why the time is ripe for discourse in machine translation.

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

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

Dario Stojanovski and Alexander Fraser. 2019. Improving anaphora resolution in neural machine translation using curriculum learning. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 140–150, Dublin, Ireland. European Association for Machine Translation.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, NIPS '17, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for 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 877–886, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, ACL '18, pages 1264–1274, Melbourne, Australia.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.

KayYen Wong, Sameen Maruf, and Gholamreza Haffari. 2020. Contextual neural machine translation improves translation of cataphoric pronouns. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5971–5978, Online. Association for Computational Linguistics.

Oksana Yakhnenko, Adrian Silvescu, and Vasant Honavar. 2005. Discriminatively trained markov model for sequence classification. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2019. Effectively training neural machine translation models with monolingual data. *Neurocomputing*, 333:240–247.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lei Li, Xin-Yu Dai, and Jiajun Chen. 2020. Mirror-generative neural machine translation. In *International Conference on Learning Representations*.

A Appendix

Model	Paramaters	Values
CONCAT	–optimizer	adam
	–adam-betas	'(0.9, 0.98)'
	–clip-norm	0.0
	–lr-scheduler	inverse_sqrt
	–warmup-init-lr	1e-07
	–warmup-updates	4000
	–lr	0.0007
	–min-lr	1e-09
	–criterion	label_smoothed_cross_entropy
	–label-smoothing	0.1
	–weight-decay	0.0
	–max-tokens	4096
	–update-freq	8
	–share-all-embeddings	-
	–max-update	100000
SEN2SEN	*as in* CONCAT	*as in* CONCAT

Table 8: Training parameters used for SEN2SEN and CONCAT models.

A.1 Training Parameters

The training parameters used for both the SEN2SEN and the CONCAT models are given in Table 8. All models were trained in **fairseq** and all results reported are based on averaging the last 10 checkpoints.

A.2 Examples from Fine-tuned Models

Some examples of improved translations from our fine-tuned models are given in Table 9.

Description	Examples
	WMT14 Testset
Source	14 stunden kämpften die ärzte um das überleben des opfers , jedoch vergeblich .
Reference	for 14 hours, doctors battled to save the life of the victim , ultimately in vain .
Baseline	14 hours of doctors fought for the victim's survival , but in vain .
Our best model	the doctors fought 14 hours for the survival of the victim , but in vain .
Source	der handel am nasdaq options market wurde am freitagnachmittag deutscher zeit unterbrochen .
Reference	trading at the nasdaq options market was interrupted on friday afternoon , german time .
Baseline	trade at nasdaq options market was cut off on the german friday afternoon .
Our best model	trade in nasdaq options market was suspended on friday afternoon in germany .
Source	einem autofahrer wurde eine strafe in höhe von 1.000 £ auferlegt , weil er mit bis zu 210 km / h und einem heißgetränk zwischen seinen beinen gefahren war .
Reference	a motorist has been fined £ 1,000 for driving at up to 130mph (210km / h) with a hot drink balanced between his legs .
Baseline	a driver was fined £ 1,000 for driving up to £ 210 per hour and a hot drink between his legs .
Our best model	a driver was fined £ 1,000 for driving up to 210 kilometers an hour and a hot drink between his legs .
Source	des grues sont arrivées sur place peu après 10 heures , et la circulation sur la nationale a été détournée dans la foulée .
Reference	cranes arrived on the site just after 10am , and traffic on the main road was diverted afterwards .
Baseline	cranes arrived soon after 10 hours , and circulation on the national front was hijacked in the process .
Our best model	cranes arrived shortly after 10 hours , and traffic on the national side was diverted along the way .
Source	le diagnostic de rage a été confirmé par l'institut pasteur .
Reference	the diagnosis of rabies was confirmed by the pasteur institute .
Baseline	the rabies diagnosis was confirmed by the institut pasteur.
Our best model	the rabies diagnosis was confirmed by the pasteur institute .
	Pronoun Testset
Context	... die die amerikanische flamme in die umnachtete welt bringe : lady liberty geht voran .
Source	sie soll die fackel der freiheit von den vereinigten staaten in den rest der welt tragen .
Context	... taking the american flame out to the benighted world : **lady liberty** is stepping forward .
Reference	she is meant to be carrying the torch of liberty from the united states to the rest of the world .
Baseline	it is meant to carry the torch of freedom from the united states to the rest of the world .
Our best model	she is supposed to carry the torch of freedom from the united states to the rest of the world .
Context	der getestete 1,6 l diesel mit 88 kw / 120 ps beschleunigt den hr - v ...
Source	er dürfte seine arbeit allerdings etwas leiser verrichten .
Context	the 1.6 l **diesel engine** we tested , with 88 kw / 120 horsepower accelerates the hr - v powerfully ...
Reference	however , it could certainly do its work a bit more quietly .
Baseline	however , he is likely to do his job rather more quietly .
Our best model	but it is likely to do its job a little more quietly .
Context	versteinerte reste der haut bedecken noch immer die holprigen panzerplatten , die den schädel des tieres tragen .
Source	sein rechter vorderfuß liegt an seiner seite , seine fünf finger sind nach oben gespreizt .
Context	fossilized remnants of skin still cover the bumpy armor plates dotting the **animal's** skull .
Reference	its right forefoot lies by its side , its five digits splayed upward .
Baseline	his right - hand front foot is on his side , his five fingers are spiked up .
Our best model	its right front foot is on its side , its five fingers are split upwards .
Context	Il est mort dimanche matin.
Source	elle avait promis à son mari , la semaine avant son décès , de le faire sortir de l'hôpital
Context	**He** died on Sunday morning.
Reference	a week before his death , she had promised her husband she would get him out of hospital
Baseline	she promised her husband , the week before she died , to take her out of the hospital .
Our best model	she promised her husband , the week before his death , to take him out of the hospital
Context	Elle a été détenue dans une cellule du commissariat local avant l'audience devant le tribunal.
Source	elle était en vacances dans la région de krabi , au sud de la thaïlande .
Context	**She** was held in local police cells before the court hearing.
Reference	she was holidaying at the resort area of krabi in southern thailand .
Baseline	it is on holiday in the region of krabi , southern thailand .
Our best model	she was on holiday in the krabi region of southern thailand .

Table 9: Examples showing the improvements in translations from our best models, across the WMT14 and the pronoun testsets. The previous sentence context information for the pronoun testset is also shown.