A Case Study on Context-Aware Neural Machine Translation with Multi-Task Learning

Ramakrishna Appicharla[1], **Baban Gain**[1], **Santanu Pal**[2], **Asif Ekbal**[3], **Pushpak Bhattacharyya**[4]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna, India
[2]Wipro AI, Lab45, London, UK
[3]School of AI and Data Science, Indian Institute of Technology Jodhpur, India
[4]Department of Computer Science and Engineering, Indian Institute of Technology Bombay, India
{ramakrishnaappicharla, gainbaban, santanu.pal.ju, asif.ekbal, pushpakbh} @gmail.com

Abstract

In document-level neural machine translation (DocNMT), multi-encoder approaches are common in encoding context and source sentences. Recent studies (Li et al., 2020) have shown that the context encoder generates noise and makes the model robust to the choice of context. This paper further investigates this observation by explicitly modelling context encoding through multi-task learning (MTL) to make the model sensitive to the choice of context. We conduct experiments on cascade MTL architecture, which consists of one encoder and two decoders. Generation of the source from the context is considered an auxiliary task, and generation of the target from the source is the main task. We experimented with German–English language pairs on News, TED, and Europarl corpora. Evaluation results show that the proposed MTL approach performs better than concatenation-based and multi-encoder DocNMT models in low-resource settings and is sensitive to the choice of context. However, we observe that the MTL models are failing to generate the source from the context. These observations align with the previous studies, and this might suggest that the available document-level parallel corpora are not context-aware, and a robust sentence-level model can outperform the context-aware models.

1 Introduction

Context-aware neural machine translation gained much attention due to the ability to incorporate context, which helps in producing more consistent translations than sentence-level models (Maruf and Haffari, 2018; Zhang et al., 2018; Bawden et al., 2018; Agrawal et al., 2018; Voita et al., 2019; Huo et al., 2020; Li et al., 2020; Donato et al., 2021). There are mainly two approaches to incorporating context. The first one is to create a context-aware input sentence by concatenating context and current input sentence (Tiedemann and Scherrer, 2017; Agrawal et al., 2018; Junczys-Dowmunt, 2019; Zhang et al., 2020b) and using it as the input to the encoder. The second approach uses an additional context-aware component to encode the source or target context (Zhang et al., 2018; Voita et al., 2018; Kim et al., 2019; Ma et al., 2020) and the entire model is jointly optimized. Typically, the current sentence's neighbouring sentences (either previous or next) are used as the context.

The context-aware models are trained to maximize the log-likelihood of the target sentence given the source sentence and context. Most of the existing works on DocNMT (Zhang et al., 2018; Maruf and Haffari, 2018; Voita et al., 2019; Li et al., 2020) focus on encoding the context through context-specific encoders. Recent studies (Li et al., 2020) show that, in the multi-encoder DocNMT models, the performance improvement is not due to specific context encoding but rather the context-encoder acts like a noise generator, which, in turn, improves the robustness of the model. In this work, we explore whether the context encoding can be modelled explicitly through multi-task learning (MTL) (Luong et al., 2015). Specifically, we aim to study the effectiveness of the MTL framework

for DocNMT rather than proposing a state-of-the-art system. The availability of document-level corpora is less compared to sentence-level corpora. Previous works (Junczys-Dowmunt, 2019) use the sentence-level corpora to warm-start the document-level model, which can be further tuned with the existing limited amount of document-level data. However, in this work, we focus only on improving the performance of DocNMT models with available document-level corpora. We consider the source reconstruction from the context as the auxiliary task and the target translation from the source as the main task. We conduct experiments on cascade MTL (Anastasopoulos and Chiang, 2018; Zhou et al., 2019) architecture. The cascade MTL architecture comprises one encoder and two decoders (Figure 1). The intermediate (first) decoder attends over the output of the encoder, and the final (second) decoder attends over the output of the intermediate decoder. The input consists of $\langle \mathrm{c_x}, \mathrm{x}, \mathrm{y} \rangle$ triplets, where $\mathrm{c_x}$, x and y represents the context, source, and target sentences, respectively. The model is trained to optimize both translation and reconstruction objectives jointly. We also train two baseline models as contrastive models, namely sentence-level vanilla baseline and single encoder-decoder model, by concatenating the context and source (Tiedemann and Scherrer, 2017; Agrawal et al., 2018; Junczys-Dowmunt, 2019). We additionally train multi-encoder single-decoder models (Li et al., 2020) to study how context affects the DocNMT models. We conduct experiments on German-English direction with three different types of contexts (*viz.* previous two source sentences, previous two target sentences, and previous-next source sentences) on News-commentary v14 and TED corpora. We report BLEU (Papineni et al., 2002) calculated with sacreBLEU (Post, 2018) and APT (accuracy of pronoun translation) (Miculicich Werlen and Popescu-Belis, 2017) scores.

To summarize, the specific attributes of our current work are as follows:

- We explore whether the MTL approach can improve the performance of context-aware NMT by introducing additional training objectives along with the main translation objective.

- We propose an MTL approach where the reconstruction of the source sentence given the context is used as an auxiliary task and the translation of the target sentence from the source sentence as the main task, jointly optimized during the training.

- The results show that in the MTL approach, the context encoder generates noise similar to the multi-encoder approach (Li et al., 2020), which makes the model robust to the choice of the context.

2 Related Work

Previous studies have proposed various document-level NMT models and achieved great success. The main goal of these approaches is to efficiently model context representation, which can lead to better translation quality. Towards this goal to represent context, Tiedemann and Scherrer (2017) concatenate consecutive sentences and use them as input to the single-encoder-based DocNMT model. Agrawal et al. (2018) conducted experiments on varying neighbouring contexts and concatenated with the current sentence as input to their model. With these similar trends, Junczys-Dowmunt (2019) conducted experiments considering the entire document as context. Further progress on context representation in DocNMT, Zhang et al. (2018) and Voita et al. (2018) proposed transformer-based multi-encoder NMT models where the additional encoder is used to encode the context. While Miculicich et al. (2018) proposed a hierarchical attention network to encode the context, a more recent approach Kang et al. (2020) proposed a reinforcement learning-based dynamic context selection module for DocNMT. Kim et al. (2019) and Li et al. (2020) conducted experiments on multi-encoder DocNMT models and reported that the performance improvement is not due to context encoding; instead, the context encoder acts as a noise generator, which improves the robustness of the DocNMT model. Junczys-Dowmunt (2019) conducted experiments on a single encoder model with masked language model objective (Devlin et al., 2019) to incorporate document-level monolingual source-side data. Since the multi-encoder models are trained to optimize the translation objective only, it might be possible for the model to pay less attention to the context, and Li et al. (2020) report the same.

MTL strategies in NMT trained on other auxiliary tasks along with the main translation task (Lu-

ong et al., 2015; Dong et al., 2015; Zaremoodi et al., 2018; Wang et al., 2020; Yang et al., 2020) achieved significant improvements in translation quality so far. The other auxiliary tasks include autoencoding (Luong et al., 2015), denoising autoencoding (Wang et al., 2020), parsing and named entity recognition (Zaremoodi and Haffari, 2018; Zaremoodi et al., 2018). Zhou et al. (2019) proposed a cascade MTL network to improve the robustness of the NMT model. They considered denoising the noisy text as an auxiliary task and the translation as the main task. They achieved a significant BLEU score improvement (up to 7.1 BLEU) on the WMT robustness shared task on the French-English dataset.

However, most multi-task models are proposed only for sentence-level NMT models. Multi-task learning is relatively unexplored in context-aware NMT settings. Wang et al. (2021) proposed an MTL framework for dialogue translation tasks that jointly correct the sentences having issues such as pronoun dropping, punctuation dropping, and typos and translate them into the target language. Liang et al. (2022) proposed a three-stage training framework for the neural chat translation task. The model is trained on auxiliary tasks such as monolingual cross-lingual response generation tasks to generate coherent translation and the next utterance discrimination task. Lei et al. (2022) proposed an MTL system to force the model to attend over relevant cohesion devices while translating the current sentence. In this work, we propose a multi-task learning objective, i.e., reconstruction of source sentences given the source context in a cascade multi-task learning setting to study the effect of context in document-level NMT systems.

3 Methodology

3.1 Problem Statement

Our document-level NMT is based on a cascade MTL framework to force the model to consider the context while generating translation. Given a source sentence x and context $\mathrm{c_x}$, the translation probability of the target sentence y in the DocNMT setting is calculated as in Equation 1.

$$p(\mathrm{y}) = p(\mathrm{y}|\mathrm{x}, \mathrm{c_x}) \times p(\mathrm{x}, \mathrm{c_x}) \tag{1}$$

We consider $p(\mathrm{x}, \mathrm{c_x})$ as the auxiliary task of source (x) reconstruction from $\mathrm{c_x}$ (as $p(\mathrm{x}|\mathrm{c_x})$) [1], calculated as in Equation 2.

$$p(\mathrm{x}, \mathrm{c_x}) = p(\mathrm{x}|\mathrm{c_x}) \times p(\mathrm{c_x}) \tag{2}$$

The training data D consists of triplets $\langle \mathrm{c_x}, \mathrm{x}, \mathrm{y} \rangle$. Given the parameters of the model θ, the translation (Equation 1) and reconstruction (Equation 2) objectives can be modeled as Equation 3 and Equation 4.

$$p(\mathrm{y}|\mathrm{x}, \mathrm{c_x}; \theta) = \prod_{t=1}^{T} p(y_t|\mathrm{x}, \mathrm{c_x}, \mathrm{y}_{<t}; \theta) \tag{3}$$

$$p(\mathrm{x}|\mathrm{c_x}; \theta) = \prod_{s=1}^{S} p(x_s|\mathrm{c_x}, \mathrm{x}_{<s}; \theta) \tag{4}$$

where, $\mathrm{S}, \mathrm{Z}, \mathrm{T}$ denote the lengths of x, $\mathrm{c_x}$, y respectively and $\mathrm{x}_{<s}, \mathrm{c_{x<z}}, y_{<t}$ denote partially generated sequences.

Given translation objective $p(\mathrm{y}|\mathrm{x}, \mathrm{c_x})$ and reconstruction objective $p(\mathrm{x}|\mathrm{c_x})$, the model is jointly trained and optimized the loss, $\mathcal{L}$ using parameter θ (cf. Equation 5); where α is a hyper-parameter used to control the loss. We set α to 0.5.

$$\begin{aligned} \mathcal{L} = \alpha * \log p(\mathrm{y}|\mathrm{x}, \mathrm{c_x}; \theta) + \\ (1 - \alpha) * \log p(\mathrm{x}|\mathrm{c_x}; \theta) \end{aligned} \tag{5}$$

We hypothesize that forcing the model to learn reconstruction and translation objectives jointly will enable the model to encode the context effectively. The output of the reconstruction task can verify this during testing. If the context encoder generates noise, then the model might be unable to reconstruct the source and vice-versa.

3.2 Cascade Multi-Task Learning Transformer

The cascade multi-task learning architecture (Zhou et al., 2019) (Figure 1) consists of one encoder and two decoders based on the transformer (Vaswani et al., 2017) architecture. The model takes three inputs: *Source*: Current source sentence, *Context*: Context of the current source sentence, and *Target*: Current target sentence. The input to the encoder is context, and the input to the intermediate decoder is the source. The intermediate decoder is trained

[1] Since the joint probability of $p(\mathrm{x}, \mathrm{c_x})$ can be calculated as $p(\mathrm{c_x}|\mathrm{x}) \times p(\mathrm{x})$, we also explored this setting. We observed that the performance of the model is poor in this setting compared to the other setting. More details can be found in Appendix A.1.

to reconstruct the source given context by attending to the output of the encoder. The final decoder attends over the output of the intermediate decoder. In the non-MTL setting, the model is trained only on the translation objective (output of the final decoder), and the intermediate decoder is not trained with the reconstruction objective.

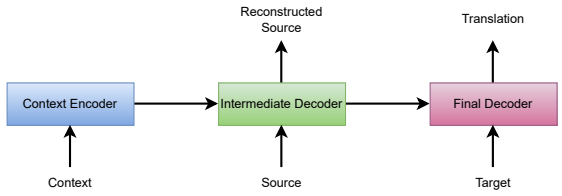

Figure 1: The overview of our MTL architecture. The input to the model is a triplet. The triplet consist of (*Context*, *Source*, *Target*). The Intermediate Decoder is trained to reconstruct the *Source* given *Context*, and the Final Decoder is trained to translate the *Source*. Here, *Source*: Current source sentence, *Context*: Context for the current source sentence, and *Target*: Translation of current source sentence. None of the layers are shared.

3.3 Context Selection

We conduct experiments on different settings of the source context. The term "source context" is defined as considering related or dependent sentences directly related to the input sentence. Based on the findings of Zhang et al. (2018), we select two sentences as context and concatenate them with a special token '$\langle \text{break} \rangle$' (Junczys-Dowmunt, 2019). For a given input source sentence (x_i) and target sentence (y_i), contexts selected for the experiments are:

- Previous-2 Source (**P@2-SRC**): Two previous source sentences ($\text{x}_{\text{i}-2}, \text{x}_{\text{i}-1}$)
- Previous-2 Target (**P@2-TGT**): Two previous target sentences ($\text{y}_{\text{i}-2}, \text{y}_{\text{i}-1}$)
- Previous-Next Source (**P-N-SRC**): Previous and next source sentences ($\text{x}_{\text{i}-1}, \text{x}_{\text{i}+1}$)

4 Experiment Setup

We train our models with the proposed cascade MTL approach. The model is trained on $\langle \text{c}_\text{x}, \text{x}, \text{y} \rangle$ triplet to jointly optimize both translation and source reconstruction objectives (Figure: 1). We also train three other contrastive models to show the effect of context in the MTL setting.

Vanilla-Sent: A vanilla sentence-level baseline model is trained without context on a single encoder-decoder network.

Concat-Context: This model is trained on a single encoder-decoder network where context is concatenated with the source (Tiedemann and Scherrer, 2017; Agrawal et al., 2018; Junczys-Dowmunt, 2019) and fed to the encoder as input. In this setting, sentences within the context are concatenated with a unique token, '$\langle \text{break} \rangle$'. The context and the source are concatenated with another special symbol, '$\langle \text{concat} \rangle$'. The special symbol helps the model to distinguish between context and source sentences.

Inside-Context: We re-implemented the 'Inside-Context' model proposed by Li et al. (2020), a multi-encoder approach. This model consists of two encoders and one decoder. The decoder is modified to include two cross-attention layers to attend over the outputs of both encoders before passing through the position-wise feed-forward layer (Vaswani et al., 2017).

4.1 Data Statistics

We conduct experiments on WMT news-commentary, IWSLT'17 TED, and Europarl-v7 German-English corpora. For the WMT news-commentary, we use news-commentary v14 (Barrault et al., 2019)[2] as the train set, newstest2017 as the validation set, and newstest2018 as the test set. For IWSLT'17 TED and Europarl-v7 corpora, we follow the train, validation, and test set splits mentioned in (Maruf et al., 2019)[3]. All models are trained on German to English. Table 1 shows data statistics of the train, validation, and test sets.

Data	# Sent	# Doc
News	329,000/3,004/2,998	8,462/130/122
TED	206,112/8,967/2,271	1,698/93/23
Europarl	1,666,904/3,587/5,134	117,855/240/360

Table 1: Data statistics for our experiments. **# Sent**, **# Doc** represent the number of sentences and documents, respectively. The numbers are shown in the Train/Validation/Test set order.

4.2 NMT Model Setups

We conduct all the experiments on transformer architecture (Vaswani et al., 2017). All the mod-

[2] https://data.statmt.org/news-commentary/v14/training/
[3] https://github.com/sameenmaruf/selective-attn/tree/master/data

Model	News		TED		Europarl	
	s-BLEU	d-BLEU	s-BLEU	d-BLEU	s-BLEU	d-BLEU
Vanilla-Sent	18.3	20.9	19.9	24.9	32.3	35.1
Concat-Context: P@2-SRC	18.0	20.5	17.3	22.4	32.5	35.4
Concat-Context: P-N-SRC	18.4	20.7	17.5	22.5	32.7	35.6
Concat-Context: P@2-TGT	14.7	17.2	15.3	20.4	**36.4**	**39.1**
MTL: P@2-SRC	19.1	21.7	20.2	24.8	29.5	32.6
MTL: P-N-SRC	**20.1**†	**22.5**	20.3	25.2	32.5†	35.3
MTL: P@2-TGT	19.2	21.7	**20.7**†	**25.4**	28.2	31.6

Table 2: BLEU scores of Vanilla-Sent, Concat-Context, and proposed MTL DocNMT models trained with different source contexts for German to English direction on News-commentary v14, IWSTL-17 TED, and Europarl corpora. **s-BLEU** and **d-BLEU** represent sentence-level and document-level BLEU respectively. The best results are shown in bold. '†' denotes the statistically significant results than Vanilla-Sent and Concat-Context models with $p < 0.05$.

els are implemented in PyTorch[4]. We use 6-layer encoder-decoder stacks with 8 attention heads. Positional token embedding sizes are set to 512, and the feed-forward layer consists of 2048 cells. Adam optimizer (Kingma and Ba, 2015) is used for training with a noam learning rate scheduler (Vaswani et al., 2017) with an initial learning rate of 0.2. We use warmup steps of 16,000 (Popel and Bojar, 2018), and dropout is set to 0.1. Due to the GPU memory restrictions, we use a mini-batch of 40 sentences for the models trained on News and TED corpora and 25 for the models trained on Europarl corpus. We create joint subword vocabularies of size 32k for each training corpus. We use the BPE (Sennrich et al., 2016) to create subword vocabularies with SentencePiece (Kudo and Richardson, 2018) implementation. We also learn the positional encoding of tokens (Devlin et al., 2019), and the maximum sequence length is set to 140 tokens for all models and 160 for *Concat-Context* models.

All the models are trained till convergence. We use the perplexity of the validation set as an early stopping criterion with the patience of 10 (Popel and Bojar, 2018). We report results on the best model checkpoint saved during the training. We perform beam search during inference with beam size 4 and length penalty of 0.6 (Wu et al., 2016). For DocNMT models, we use the same source context with which the models are trained. Since the input to the intermediate decoder (source sentence) is also given during the testing phase, the representation of the intermediate decoder can be calculated in parallel, similar to the training phase.

All the experiments are conducted on a single Nvidia GTX 2080ti GPU. The number of parameters and training time of the models is as follows: *Vanilla-Sent*: 76M, 76.5 hours, *Concat-Context*: 76M, 81 hours, *Inside-Context*: 118M, 125 hours and proposed *MTL*: 130M, 160 hours. The parameters and training times are approximately the same for all the corpora.

[4]https://pytorch.org/

5 Results and Analysis

This section discusses the results of the trained models and the context's effect on Multi-Encoder and MTL settings. Table 2 shows the sentence-BLEU (s-BLEU) and document-BLEU (d-BLEU) (Liu et al., 2020; Bao et al., 2021) scores of the proposed multi-task learning model along with the *Vanilla-Sent* and *Concat-Context* models.

We report all models' BLEU scores on German $\rightarrow$ English direction, calculated with sacreBLEU (Post, 2018).

5.1 Results of MTL and Contrastive Models

We report the BLEU scores of the models on German $\rightarrow$ English direction, calculated with sacreBLEU (Post, 2018)[5]. The proposed MTL model can outperform both *Vanilla-Sent* and *Concat-Context* models by achieving s-BLEU scores of 20.1 (*MTL: P-N-SRC*) and 20.7 (*MTL: P@2-TGT*) with an improvement of +1.8 and +0.8 BLEU improvement for News and TED corpora respectively. However, in the case of the Europarl data set, *Concat-Context* models outperform both *Vanilla-Sent* and *MTL* models. This shows that the *Concat-Context* model requires more data to perform well, unlike the MTL models, which can also work effectively in low-resource settings. We observe that the performance of the models is almost uniform across the three different context settings

[5]sacreBLEU signature:"nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1"

with a maximum BLEU difference of +1.0 (*P-N-SRC* vs. *P@2-SRC*) on News, +0.5 (*P@2-TGT* vs. *P@2-SRC*) on TED and +4.3 (*P-N-SRC* vs *P@2-TGT*) on Europarl corpora respectively.

We also report d-BLEU (document-level BLEU) scores (Liu et al., 2020; Bao et al., 2021) by converting each document into one single sequence (paragraph) by concatenating all sentences from that document and calculate BLEU scores on the resulting corpus. This results in slightly higher scores than the sentence level by matching n-grams over the whole document instead of at the sentence level. Table 2 also shows d-BLEU scores. Like s-BLEU scores, proposed MTL models achieve the best d-BLEU scores of 22.5 and 25.4 for News and TED corpora, respectively. We report the paired bootstrap resampling (Koehn, 2004) results, calculated with sacreBLEU (Post, 2018).

Model	News	TED	Europarl
MTL: P@2-SRC	1.3	1.4	4.9
MTL: P@2-TGT	1.2	1.6	3.9
MTL: P-N-SRC	1.3	1.5	3.1

Table 3: s-BLEU scores for the reconstruction objective of the MTL models on test set for News, TED, and Europarl corpora.

5.2 Analysis of Reconstruction Objective

We analyze the performance of the MTL model on the reconstruction objective on the test set to verify if the context encoder is generating noise. If the context encoder generates noise by the suboptimal encoding of context, the intermediate decoder will fail to reconstruct the source sentence from the context; otherwise, the intermediate decoder can reconstruct the source sentence to a similar extent as the final translated sentence. We perform greedy decoding on the intermediate decoder to generate the source from the context. Table 3 shows the BLEU scores of the reconstruction objective on the test set for News, TED, and Europarl corpora. The results show that the MTL models fail to reconstruct the source from the context. Based on this, we conclude that the context encoder cannot encode the context, leading to poor reconstruction performance of the models. However, we hypothesize that the model cannot reconstruct the source from the context because the corpora used to train context-aware models might not be context-aware. This observation aligns with the previous works (Kim et al., 2019; Li et al., 2020), and with enough data, vanilla sentence-level NMT models can outperform the document-level NMT models.

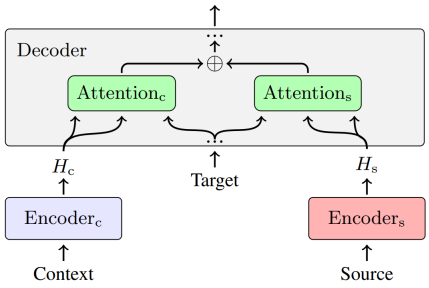

Figure 2: The overview of the Inside-Context model. The input to the model is a triplet consisting of (*Context, Source, Target*). The multi-head attention layer of the decoder is modified to attend to both the context encoders (Encoder_c) and the source encoder (Encoder_s).

Model	News	TED	Europarl
MTL: P@2-SRC	19.1	20.2	29.5
MTL: P-N-SRC	**20.1**†	20.3	32.5†
MTL: P@2-TGT	19.2	**20.7**†	28.2
Inside-Context: P@2-SRC	18.8	19.6	33.2
Inside-Context: P-N-SRC	19.0	19.8	33.2
Inside-Context: P@2-TGT	18.3	20.4	**33.6**

Table 4: Comparison of s-BLEU scores of MTL and Inside-Context Multi-Encoder models. The best results are shown in bold. '†' denotes the statistically significant results than Vanilla-Sent and Concat-Context models with $p < 0.05$.

5.3 MTL vs. Multi-Encoder Approach

We compare the proposed MTL approach to the existing Multi-Encoder approach to study how the model will perform in a single-task setting. Specifically, we compare our MTL approach (single-encoder multi-decoder network) with *Inside-Context* (Li et al., 2020) architecture. This model consists of two transformer encoders and one transformer decoder. Figure 2 shows the model's architecture. The decoder is modified to attend to the outputs of both encoders. The model follows the transformer (Vaswani et al., 2017) architecture. An element-wise addition is performed on the outputs of both cross-attention layers before passing through layer-norm and position-wise feed-forward layers. Table 4 shows the s-BLEU scores of the MTL and Inside-Context models. We observe that the performance of multi-encoder models is similar to MTL models, with MTL models achieving +1.1 (P-N-SRC models), +0.3 (P@2-TGT models) BLEU points im-

provement over Inside-Context models for News and TED corpora respectively. In the case of Europarl, inside-context models achieve better performance than the MTL models, with the P@2-TGT model achieving +5.4 BLEU points improvement compared to the MTL model. Based on the results, we conclude that the MTL setting is more effective for low-resource scenarios.

Model	News	TED	Europarl
MTL: P@2-SRC	1.2 (-17.9)	0.8 (-19.4)	4.5 (-25.0)
MTL: P-N-SRC	1.2 (-18.9)	0.8 (-19.5)	4.0 (-28.5)
MTL: P@2-TGT	0.5 (-18.7)	0.3 (-20.4)	3.9 (-24.3)
Inside-Context: P@2-SRC	18.7 (-0.1)	19.4 (-0.2)	33.2 (0.0)
Inside-Context: P-N-SRC	18.9 (-0.1)	19.8 (0.0)	33.2 (0.0)
Inside-Context: P@2-TGT	18.3 (0.0)	20.3 (-0.1)	33.1 (-0.5)

Table 5: Comparison of s-BLEU scores of MTL models tested with random context. The difference in scores over the models trained with the selected context is shown inside the parentheses.

5.4 Effect of Context in MTL setting

Since the BLEU scores of our MTL models are almost the same for all three context settings, we check whether the MTL models are affected by the choice of context. To this end, we test the MTL models with random context. Here, random context denotes two randomly selected sentences from the entire corpus. Table 5 shows the results of MTL and Inside-Context models tested with random context. Results show that the MTL models fail to translate source sentences when the context is random. However, Inside-Context models are agnostic to context as models can translate well even if the context is random. Our findings in the case of multi-encoder models are in line with the findings of Li et al. (Li et al., 2020). Based on the results, we conclude that MTL models are sensitive to the choice of context. Section A.1.1 describes a similar experiment where the MTL models are tested with random context. However, the architecture used in the main experiments differs slightly from the one used in the preliminary investigation. We observe that feeding the Intermediate Decoder output to the Final Decoder makes the model sensitive to the choice of context (cf. Figure 1 and Figure 3 in the Appendix A.1). We hypothesize that a weighted combination of the Context Encoder output and Intermediate Decoder output is desired as it performs slightly better than the model used in the main experimental setup. However, it also makes the model agnostic to the choice of context. We plan to explore this behaviour in detail in our future work.

Model	News	TED	Europarl
MTL: P@2-SRC	13.7 (+12.5)	11.2 (+10.4)	22.3 (+17.8)
MTL: P-N-SRC	14.5 (+13.3)	11.3 (+10.5)	19.7 (+15.7)
Inside-Context: P@2-SRC	18.7 (0.0)	19.6 (+0.2)	33.1 (-0.1)
Inside-Context: P-N-SRC	19.0 (+0.1)	19.7 (-0.1)	33.0 (-0.2)

Table 6: s-BLEU scores of the MTL and Inside-Context models are tested by giving the same source sentences as context and input. The change of s-BLEU scores over the models tested with random context is shown in ($\pm x$).

5.5 Results of MTL and Multi-Encoder models without Context

We conduct experiments on MTL and Inside-Context models by using the same source sentence as the context. Since the proposed MTL models fail when tested with random context (cf. Section 5.4), we observe how the MTL and Multi-Encoder models are performing when the same source sentence is given as context. This setting presents a scenario where the context is not random but also not the type of context with which the models are trained. We conduct experiments for *P@2-SRC* and *P-N-SRC* context settings only as the *P@2-TGT* context setting requires the current target sentence, which is unavailable during testing. We observe that MTL models can perform well compared to the random context setting, which shows that the MTL models are sensitive to the choice of context. The performance of Inside-Context models is almost the same as those tested with random context. This shows that the Inside-Context model is agnostic to the choice of the context. Table 6 shows the s-BLEU scores of the MTL and Inside-Context models.

Model	News	TED	Europarl
Vanilla-Sent	40.17	31.22	37.22
Concat-Context: P@2-SRC	39.34	30.01	36.42
Concat-Context: P-N-SRC	39.99	29.57	36.78
Concat-Context: P@2-TGT	38.50	28.82	**37.27**
MTL: P@2-SRC	40.69	31.44	35.96
MTL: P-N-SRC	40.50	31.24	36.94
MTL: P@2-TGT	**40.99**	**31.90**	33.91

Table 7: Accuracy of Pronoun Translation (APT) scores. The best results are shown in bold.

5.6 Pronoun Translation Accuracy

We also evaluate our proposed models' performance on pronoun translation accuracy. We

calculate the pronoun translation accuracy with APT (accuracy of pronoun translation) (Miculicich Werlen and Popescu-Belis, 2017) metric[6]. This metric requires a list of pronouns from the source language (German) with a list of pronouns from the target language (English) as an optional argument. We use spaCy[7] to tag both source and target sentences from the test set and extract pronouns. Table 7 shows the APT scores of *Vanilla-Sent*, *Concat-Context*, and *MTL* DocNMT models. The APT scores correlate with the s-BLEU and d-BLEU scores, achieving the highest APT score of 40.99 in *MTL: P@2-TGT* setting with an improvement of +0.82 over *Vanilla-Sent* and +1.0 over *Concat-Context* (*P-N-SRC*) models on News corpus. Similarly, the *MTL: P@2-TGT* model achieves the highest APT score of 31.90 with an improvement of +0.68 and +1.89 over *Vanilla-Sent* and *Concat-Context* (*P@2-SRC*) on TED. For the Europarl corpus, *Concat-Context* (*P@2-TGT*) achieved the highest APT score of 37.27 with an improvement of +0.05 and +0.33 over *Vanilla-Sent* and *MTL* (*P-N-SRC*) models respectively.

6 Conclusion

This work explored the MTL approach for document-level NMT (DocNMT). Our proposed MTL approach is based on cascade MTL architecture, where the model consists of one encoder (for context encoding) and two decoders (for the representation of the current source and target sentences). Reconstruction of the source sentence given the context is considered the auxiliary task, along with the translation of the current source sentence as the main task. We conducted experiments for German–English for News-commentary v14, IWSLT'17 TED, and Europarl v7 corpora with three different types of contexts *viz.* two previous sources, two previous targets, and previous-next source sentences with respect to the current input source sentence.

Our proposed MTL approaches outperform the sentence-level baseline and concatenated-context models in low-resource (for News and TED corpora) settings. However, all models perform well in the high resource setting (Europarl corpus), with proposed MTL models slightly underperforming the rest. Our MTL models are more sensitive to the choice of context than the multi-encoder models when tested with random context. We observe that the context encoder cannot encode context sufficiently and performs poorly reconstruction tasks. Finally, we reported APT (accuracy of pronoun translation) scores, and the proposed MTL models outperformed the sentence-level baseline and concatenated-context models. Our empirical analysis concludes that our approach is more sensitive to the choice of context and improves the overall translation performance in low-resource context-aware settings. We plan to explore other tasks, such as gap sentence generation (GSG) (Zhang et al., 2020a) as an auxiliary task for better context encoding, different training curricula to prioritize one objective over the other during the training, and dynamic context selection.

7 Limitations

Our study poses two main limitations. First, our primary motivation is to understand the effect of context and if the context encoding can be modelled as an auxiliary task but not to propose a model to achieve state-of-the-art results. We have followed the findings of Li et al. (Li et al., 2020) and used one of their approach to understanding the effect of context. Our observations are also in line with their findings.

Second, even though our proposed MTL approach can outperform the models in other settings, the auxiliary task (reconstruction) is not very effective as it improves the BLEU scores in the range of [0.1-1.8] over the Multi-encoder models. We hypothesize that, in the loss function, we are giving equal weights to both the objectives (0.5 for both reconstruction and translation objectives), which might lead to significantly less improvement in overall translation quality. We plan to explore different training curricula to adjust the weight of the objectives dynamically during the training.

Acknowledgements

We gratefully acknowledge the support from the "NLTM: VIDYAAPATI" project, sponsored by Electronics and IT, Ministry of Electronics and Information Technology (MeiTY), Government of India. Santanu Pal acknowledges the support from Wipro AI. We also thank the anonymous reviewers for their insightful comments.

[6] https://github.com/idiap/APT
[7] https://spacy.io/models

References

Agrawal, Ruchit Rajeshkumar, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *21st Annual Conference of the European Association for Machine Translation*, pages 11–20.

Anastasopoulos, Antonios and David Chiang. 2018. Tied multitask learning for neural speech translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana, June. Association for Computational Linguistics.

Bao, Guangsheng, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3442–3455, Online, August. Association for Computational Linguistics.

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

Bawden, Rachel, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana, June. Association for Computational Linguistics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Donato, Domenic, Lei Yu, and Chris Dyer. 2021. Diverse pretrained context encodings improve document translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1299–1311, Online, August. Association for Computational Linguistics.

Dong, Daxiang, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China, July. Association for Computational Linguistics.

Huo, Jingjing, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 604–616, Online, November. Association for Computational Linguistics.

Junczys-Dowmunt, Marcin. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy, August. Association for Computational Linguistics.

Kang, Xiaomian, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online, November. Association for Computational Linguistics.

Kim, Yunsu, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34, Hong Kong, China, November. Association for Computational Linguistics.

Kingma, Diederik P. and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Bengio, Yoshua and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Kudo, Taku and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Kudo, Taku. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, July. Association for Computational Linguistics.

Lei, Yikun, Yuqi Ren, and Deyi Xiong. 2022. CoDoNMT: Modeling cohesion devices for document-level neural machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5205–5216, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Li, Bei, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online, July. Association for Computational Linguistics.

Liang, Yunlong, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022. Scheduled multi-task learning for neural chat translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4375–4388, Dublin, Ireland, May. Association for Computational Linguistics.

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Luong, Minh-Thang, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Ma, Shuming, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online, July. Association for Computational Linguistics.

Maruf, Sameen and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1275–1284, Melbourne, Australia, July. Association for Computational Linguistics.

Maruf, Sameen, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Miculicich, Lesly, 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, October-November. Association for Computational Linguistics.

Miculicich Werlen, Lesly and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark, September. Association for Computational Linguistics.

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

Popel, Martin and Ondřej Bojar. 2018. Training tips for the transformer model. *arXiv preprint arXiv:1804.00247*.

Post, Matt. 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, October. Association for Computational Linguistics.

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

Tiedemann, Jörg 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, September. Association for Computational Linguistics.

Vaswani, Ashish, 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.

Voita, Elena, 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 (Volume 1: Long Papers), pages 1264–1274, Melbourne, Australia, July. Association for Computational Linguistics.

Voita, Elena, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, July. Association for Computational Linguistics.

Wang, Yiren, ChengXiang Zhai, and Hany Hassan. 2020. Multi-task learning for multilingual neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1022–1034, Online, November. Association for Computational Linguistics.

Wang, Tao, Chengqi Zhao, Mingxuan Wang, Lei Li, and Deyi Xiong. 2021. Autocorrect in the process of translation — multi-task learning improves dialogue machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 105–112, Online, June. Association for Computational Linguistics.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Yang, Jiacheng, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. Towards making the most of bert in neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9378–9385.

Zaremoodi, Poorya and Gholamreza Haffari. 2018. Neural machine translation for bilingually scarce scenarios: a deep multi-task learning approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1356–1365, New Orleans, Louisiana, June. Association for Computational Linguistics.

Zaremoodi, Poorya, Wray Buntine, and Gholamreza Haffari. 2018. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 656–661, Melbourne, Australia, July. Association for Computational Linguistics.

Zhang, Jiacheng, 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, October-November. Association for Computational Linguistics.

Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Zhang, Pei, Boxing Chen, Niyu Ge, and Kai Fan. 2020b. Long-short term masking transformer: A simple but effective baseline for document-level neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1081–1087, Online, November. Association for Computational Linguistics.

Zhou, Shuyan, Xiangkai Zeng, Yingqi Zhou, Antonios Anastasopoulos, and Graham Neubig. 2019. Improving robustness of neural machine translation with multi-task learning. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 565–571, Florence, Italy, August. Association for Computational Linguistics.

A Appendix

A.1 Preliminary Investigation on Auxiliary Objectives

The joint probability in Equation 2 ($p(\mathrm{x}, \mathrm{c_x})$) can be calculated in two ways such as:

$$p(\mathrm{x}, \mathrm{c_x}) = p(\mathrm{x}|\mathrm{c_x}) \times p(\mathrm{c_x}) \quad (6)$$

$$p(\mathrm{x}, \mathrm{c_x}) = p(\mathrm{c_x}|\mathrm{x}) \times p(\mathrm{x}) \quad (7)$$

Since the joint probability can be computed in two different ways, we conduct an initial study to select the optimal auxiliary objective that improves the overall translation performance of the model. Specifically, we consider $p(\mathrm{x}|\mathrm{c_x})$ as one auxiliary task where source (x) is autoregressively reconstructed (denoted as **Re-Src**) from the encoded context ($\mathrm{c_x}$) and $p(\mathrm{c_x}|\mathrm{x})$ as the other auxiliary task where context (x) is autoregressively

reconstructed (denoted as **Re-Cntx**) from the encoded source (c_x. We conducted experiments to verify which auxiliary task is performing better.

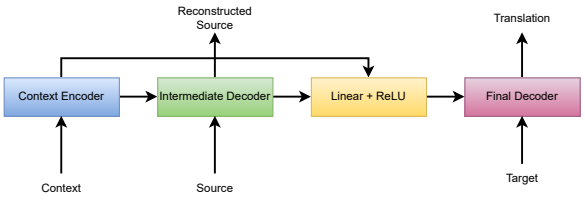

Figure 3: The overview of modified MTL architecture with residual connection. The input to the model is a triplet. The triplet consist of (*Context*, *Source*, *Target*) in **Re-Src** setting and (*Source*, *Context*, *Target*) in **Re-Cntx** setting. Here, *Source*: Current source sentence, *Context*: Context for the current source sentence, and *Target*: Translation of current source sentence. None of the layers are shared.

The experimental setup and model architecture are slightly different for this comparison study than those used in the main experiments.[8] The Context Encoder and Intermediate Decoder output are combined with a linear layer with ReLU activation. The main experimental setup does not use this linear layer + ReLU combination. We hypothesize that adding this layer might make the model agnostic to the choice of context. We test this by training the model with random context (cf. Section A.1.1. Specifically, we use two context settings *viz. P@2-SRC* and *P-N-SRC* settings (cf. 3.3). We use a fixed learning rate of 10^{-5} instead of the warmup schedule. The output from this layer is given as input to the Final Decoder.

Model		Vanilla-Sent	MTL: P@2-SRC	MTL: P-N-SRC
News	Re-Src	16.5	20.6	**20.9**
	Re-Cntx		16.7 (-3.9)	17.9 (-3.0)
TED	Re-Src	12.1	21.6	**22.0**
	Re-Cntx		18.0 (-3.6)	17.8 (-4.2)
Europarl	Re-Src	35.0	35.1	**35.8**
	Re-Cntx		33.2 (-1.9)	33.6 (-2.2)

Table 8: Comparison of s-BLEU scores of Baseline and proposed MTL DocNMT models trained with different source contexts for German to English direction. Differences in the scores over **Re-Src** are shown inside the parentheses.

We use a mini-batch of 18 sentences to train all the models. We create two separate subword vocabularies for each training corpus. The created subword vocabulary is 40k in both German and English. We use the unigram language model (Kudo, 2018) to create subword vocabularies with SentencePiece (Kudo and Richardson, 2018), and the maximum sequence length is set to 160 tokens. During inference, we perform greedy decoding. The rest of the experimental setup is the same as the one used in the main experiments.

[8]We modified the experimental setup and model architecture during our main experiments. In this preliminary investigation, the capacity of models with independent subword vocabularies is slightly larger. Due to this, the s-BLEU scores are slightly better than the main results.

Model	Random-Train		Random-Infer	
	Re-Src	Re-Cntx	Re-Src	Re-Cntx
MTL: P@2-SRC	20.9	16.6	20.6	16.8
MTL: P-N-SRC	20.9	16.4	20.8	17.8

Table 9: s-BLEU scores of *Random-Train* and *Random-Infer* experiments on News-commentary corpus.

A.1.1 Effect of Random Context

We also conduct experiments to study how the random context affects the MTL models. Specifically, we evaluate the MTL models in two settings. The model is trained on the random context in the *Random-Train* setting by concatenating two randomly sampled sentences from the train set and testing with *P@2-SRC* and *P-N-SRC* context settings. In *Random-Infer* setting, the model is trained on *P@2-SRC* and *P-N-SRC* context settings and tested with random context. We train the models on the news-commentary corpus. Table 9 shows the s-BLEU scores of the MTL models trained and tested in the random context setting. Based on the results, we conclude that the model trained with random context improves the robustness of the model. This observation aligns with the findings of Li et al. (2020), but they conducted experiments in the non-MTL setting with multiple encoders. As the model largely ignores the choice of the context, we remove this linear + ReLU combination and feed the output of the Intermediate Decoder to the Final Decoder. We hypothesize that this forces the model to consider the context while generating the target sentence.