# Importance-Aware Data Augmentation for Document-Level Neural Machine Translation

**Minghao Wu**[♡] **Yufei Wang**[♡] **George Foster**[♠] **Lizhen Qu**[♡] **Gholamreza Haffari**[♡]
[♡]Monash University [♠]Google Research
{firstname.lastname}@monash.edu fosterg@google.com

## Abstract

Document-level neural machine translation (DocNMT) aims to generate translations that are both coherent and cohesive, in contrast to its sentence-level counterpart. However, due to its longer input length and limited availability of training data, DocNMT often faces the challenge of data sparsity. To overcome this issue, we propose a novel **I**mportance-**A**ware **D**ata **A**ugmentation (IADA) algorithm for DocNMT that augments the training data based on token importance information estimated by the norm of hidden states and training gradients. We conduct comprehensive experiments on three widely-used DocNMT benchmarks. Our empirical results show that our proposed IADA outperforms strong DocNMT baselines as well as several data augmentation approaches, with statistical significance on both sentence-level and document-level BLEU.

**Context**

~~**Because**~~ **of paralysis, my grandmother's legs** ~~**have**~~ **stopped working.**

**Current Sentence**

**Today,** ~~**she**~~ **had another** ~~**attack**~~**.**

Figure 1: An example showing the missing information can be recovered by the complementary information in the context. ~~**Strikethrough**~~ indicates perturbation.

## 1 Introduction

Document-level Neural Machine Translation (DocNMT) has achieved significant progress in recent years, as evidenced by notable studies (Tiedemann and Scherrer, 2017; Maruf and Haffari, 2018; Wong et al., 2020; Wu et al., 2021; Li et al., 2022; Lupo et al., 2022; Sun et al., 2022; Wang et al., 2023; Lyu et al., 2023; Wu et al., 2024). By effectively incorporating contextual information, DocNMT aims to enhance the coherence and cohesion between the translated sentences, compared with its sentence-level counterpart (SentNMT). However, training DocNMT models requires document-level parallel corpora, which are more difficult and expensive to obtain than SentNMT. This data sparsity issue can cause DocNMT models to learn spurious patterns in the training data, leading to poor generalization (Dankers et al., 2022).

To overcome this issue, the *data augmentation* (DA) technology (Shorten et al., 2021; Wang et al., 2022) offers a promising solution. These DA methods for SentNMT typically generate synthetic data by randomly perturbing tokens in the training instances (Gal and Ghahramani, 2016; Sennrich et al., 2016a; Wei and Zou, 2019; Takase and Kiyono, 2021). On top of this, in this paper, we propose a novel **I**mportant-**A**ware **D**ata **A**ugmentation (IADA) method, which provides explicit signals for training the DocNMT models to proactively utilize document contextual information. Specifically, as shown in Figure 1, IADA first perturbs the *important* tokens (i.e., *she* and *attack*) in the current sentence to be translated, which enforces the DocNMT models to recover those information using the document context. IADA further perturbs the *less important* tokens in the context (i.e., *because* and *have*), highlighting the useful information in the document context. To determine *token importance*, we propose two novel measures derived from the DocNMT model: the topmost hidden states of the encoder/decoder (TNorm), which leverages context-dependent information, and training gradients (GNorm), which takes source-target alignment information into account. Finally, as IADA perturbs the important information in current sentences and could increase learning difficulty. We combat this issue by adding an agreement loss between the original and perturbed instances.

In this work, we combine IADA with two pop-



*Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 740–752
March 17-22, 2024 

ular data augmentation methods, word dropout (Gal and Ghahramani, 2016) (i.e., IADA$_{\text{DROP}}$) and word replacement (Takase and Kiyono, 2021) (i.e., IADA$_{\text{REPL}}$). We evaluate these versions on three widely-used DocNMT benchmarks: `TED`, `News`, and `Europarl`. Our experiments consistently demonstrate that both IADA$_{\text{DROP}}$ and IADA$_{\text{REPL}}$ outperform various strong Doc2Doc models with statistical significance. We perform ablation studies to validate the effectiveness of our design choices. Through our analyses, we show that IADA enhances contextual awareness and robustness in the DocNMT model. Additionally, we demonstrate that IADA can be combined with back/forward-translation techniques and is particularly beneficial in low-resource settings. Lastly, our linguistic study confirms IADA's ability to effectively identify important tokens in the text.

## 2 Related Work

**Document-Level NMT** In recent years, numerous approaches have been proposed for document-level neural machine translation (DocNMT). One early model, proposed by Tiedemann and Scherrer (2017), simply concatenates the context and the current sentence. Since then, many works on DocNMT have been published, covering various research topics such as model architecture (Miculicich et al., 2018; Maruf et al., 2019; Zhang et al., 2021; Wu et al., 2023), training methods (Sun et al., 2022; Lei et al., 2022), and evaluation (Bawden et al., 2018; Jiang et al., 2022). Unlike its sentence-level NMT (SentNMT), DocNMT often faces data scarcity issues, as collecting parallel document pairs is even more challenging and expensive, impeding the progress of DocNMT.

**Data Augmentation** Data augmentation (DA) approaches for NMT are commonly categorized into two classes, word replacement and back/forward translation. Gal and Ghahramani (2016) and Sennrich et al. (2016a) introduce word dropout (WordDrop), where word embeddings are zeroed out at random positions in the input sequence. Provilkov et al. (2020) incorporate a dropout-like mechanism into the BPE segmentation process (Sennrich et al., 2016c; Kudo, 2018), generating multiple segments for the same sequence. Liu et al. (2021) utilize language models and phrasal alignment with causal modeling to augment sentence pairs. Takase and Kiyono (2021) demonstrate that word dropout (WordDrop) and word replacement (WordRepl) can achieve strong performance with improved computational efficiency. Kambhatla et al. (2022) expand the training corpus by enciphering the text with deterministic rules. Back-translation (BT) translates the monolingual corpus from the target language back to the source language, resulting in significant performance improvements (Bojar and Tamchyna, 2011; Sennrich et al., 2016b). Hoang et al. (2018) perform iterative BT and observe substantial performance gains. Another approach, known as forward-translation (FT) or self-training, translates the monolingual source corpus into the target language (Zhang and Zong, 2016; He et al., 2020). Recent works perform BT with a DocNMT model, known as DocBT (Huo et al., 2020; Ul Haq et al., 2020).

**Ours** Our novel Important-Aware Data Augmentation (IADA) method effectively encourages the DocNMT model to leverage the contextual information. Our empirical results conform that IADA is compatible with the classical DA approaches, such as DocBT and DocFT.

## 3 Method

In this section, we introduce the task of DocNMT in Section 3.1, our proposed IADA framework in Section 3.2, our *token importance measures* in Section 3.3, and our training objective in Section 3.4.

### 3.1 Document-Level NMT

The standard sentence-level NMT (SentNMT) model ignores surrounding context information, whose probability of translation is defined as:

$$P(\boldsymbol{y}_i|\boldsymbol{x}_i) = \prod_{t=1}^{|\boldsymbol{y}_i|} P(y_{i,t}|\boldsymbol{y}_{i,<t}, \boldsymbol{x}_i), \quad (1)$$

where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are the $i$-th source and target training sentence, $y_{i,t}$ denotes the $t$-th token in $\boldsymbol{y}_i$ and $|\cdot|$ indicates the sequence length. Different from SentNMT, DocNMT has the access to both current sentence and context sentences for translation. Given a document pair $\{\boldsymbol{X}_i, \boldsymbol{Y}_i\}$, we define $\boldsymbol{X}_i = \{\boldsymbol{C}_{\boldsymbol{x}_i}, \boldsymbol{x}_i\}$ and $\boldsymbol{Y}_i = \{\boldsymbol{C}_{\boldsymbol{y}_i}, \boldsymbol{y}_i\}$, where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are the current sentence pair, and $\boldsymbol{C}_{\boldsymbol{x}_i}$ and $\boldsymbol{C}_{\boldsymbol{y}_i}$ are their corresponding context. The translation probability of $\boldsymbol{y}_i$ in DocNMT is:

$$P(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{C}_{\boldsymbol{x}_i}, \boldsymbol{C}_{\boldsymbol{y}_i}) = \prod_{t=1}^{|\boldsymbol{y}_i|} P(y_{i,t}|\boldsymbol{y}_{i,<t}, \boldsymbol{x}_i, \boldsymbol{C}_{\boldsymbol{x}_i}, \boldsymbol{C}_{\boldsymbol{y}_i}), \quad (2)$$

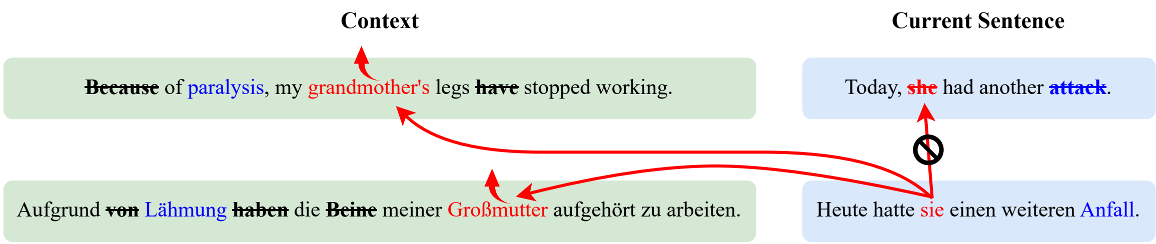


Figure 2: An illustrative example of IADA. ~~Strikethrough~~ indicates perturbation. The "sie" is semantically connected to "she", "grandmother", and "Großmutter". IADA is inclined to mask "she" in the current sentence and other less-important words in the context. Tokens in blue are similarly affected by IADA.

### 3.2 Importance-Aware Data Augmentation

Existing DocNMT models only demonstrate limited usage of the context (Fernandes et al., 2021), while an ideal one should proactively leverage the contextual information in the translation process. **I**mportance-**A**ware **D**ata **A**ugmentation (IADA) is built on top of this goal. Specifically, IADA first perturbs the *important* tokens in the current sentence to be translated, which encourages the DocNMT models to recover those information using the document context. IADA then perturbs the *less important* tokens in the context, highlighting the useful contextual information. Note that these two steps can be performed simultaneously.

As shown in Figure 2, IADA is likely to perturb "she" and "attack" in the current sentence and "because" and "have" in the context. Accordingly, after IADA perturbation, the context sentences generally have more valuable information than the current sentences, providing the inductive bias that context is crucial during training.

To implement this design, IADA perturbs the original document pair and obtain $\tilde{\boldsymbol{X}}_i = \{\tilde{\boldsymbol{C}}_{\boldsymbol{x}_i}, \tilde{\boldsymbol{x}}_i\}$ and $\tilde{\boldsymbol{Y}}_i = \{\tilde{\boldsymbol{C}}_{\boldsymbol{y}_i}, \tilde{\boldsymbol{y}}_i\}$. Accordingly, the translation probability of a DocNMT model with IADA is:

$$P(\boldsymbol{y}_i|\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{C}}_{\boldsymbol{x}_i}, \tilde{\boldsymbol{C}}_{\boldsymbol{y}_i}) = \prod_{t=1}^{|\boldsymbol{y}_i|} P(y_{i,t}|\tilde{\boldsymbol{y}}_{i,<t}, \tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{C}}_{\boldsymbol{x}_i}, \tilde{\boldsymbol{C}}_{\boldsymbol{y}_i}), \tag{3}$$

IADA uses a *token-specific replacement probability* $p_{i,t}$ to determine the tokens to be replaced in these sentences. For example, the token $x_{i,t}$ in the source document $\boldsymbol{X}_i$ is replaced:

$$\begin{aligned} m_{i,t} &\sim \text{Bernoulli}(p_{i,t}), \\ \tilde{x}_{i,t} &= \begin{cases} \Omega(x_{i,t}), & \text{if } m_{i,t} = 1; \\ x_{i,t}, & \text{otherwise}, \end{cases} \end{aligned} \tag{4}$$

where $\Omega(\cdot)$ could be an arbitrary replacement strategy. IADA can be incorporated with various existing replacement strategies. In this paper, we show the effectiveness of two versions of IADA, IADA$_{\text{DROP}}$ (with word dropout) and IADA$_{\text{REPL}}$ (with word replacement).

**Token-Specific Replacement Probability** As discussed above, in IADA, the important tokens in the context should be assigned lower replacement probabilities, while the important tokens in the current sentence should be assigned higher replacement probabilities. Therefore, for the token $x_{i,t}$ in the source document $\boldsymbol{X}_i$, we define its corresponding $p_{i,t}$ as:

$$p_{i,t} = \begin{cases} \sigma(\sigma^{-1}(p_{\text{ctx}}) - \psi(x_{i,t})), \text{if } x_{i,t} \in \boldsymbol{C}_{\boldsymbol{x}_i}, \\ \sigma(\sigma^{-1}(p_{\text{cur}}) + \psi(x_{i,t})), \text{if } x_{i,t} \in \boldsymbol{x}_i, \end{cases} \tag{5}$$

where $p_{\text{ctx}}$ and $p_{\text{cur}}$ are the initial replacement probabilities for the context and current sentence respectively, and $\sigma(\cdot)$ is the sigmoid function whose output can be interpreted as a probability.

**Importance Normalization** To properly control the spread of token importance scores, we propose to normalize the token importance score $\psi(x_{i,t})$ across all tokens in $\boldsymbol{X}_i$ as:

$$\psi(x_{i,t}) = \alpha \frac{\phi(x_{i,t}) - \mu_i}{\sigma_i}, \tag{6}$$

where

$$\mu_i = \frac{1}{|\boldsymbol{X}_i|} \sum_{t=1}^{|\boldsymbol{X}_i|} \phi(x_{i,t}), \tag{7}$$

$$\sigma_i = \sqrt{\frac{1}{|\boldsymbol{X}_i|} \sum_{t=1}^{|\boldsymbol{X}_i|} (\phi(x_{i,t}) - \mu_i)^2}. \tag{8}$$

$\phi(x_{i,t})$ is the original token importance score. $\alpha$ is the hyper-parameter that controls the spread of token importance scores. We also apply this normalization process to $\psi(y_{i,t})$ in the target documents.

### 3.3 Token Importance Measures

In this section, we discuss how IADA determines the word importance score $\phi(x_{i,t})$ for the DocNMT training instances. Schakel and Wilson (2015) and Wilson and Schakel (2015) discover that words only used in specific context are often associated with higher values of word embedding norm. These words often refer to the concrete real world objects/concepts and should be considered as important words in the sentence (Luhn, 1958). Motivated by these findings, we propose two different approaches to leverage the internal states of input tokens in the DocNMT models in $\phi(x_{i,t})$.

**Norm of Topmost Hidden States (TNorm)** The meaning of a word is dynamic according to its surrounding context. Thus, we propose to use the norm of topmost layer hidden states $\boldsymbol{h}_{x_{i,t}}$ from encoder, which incorporates the context-aware information (Peters et al., 2018; Devlin et al., 2019), as importance measure. The importance measure $\phi_{\text{TNorm}}(x_{i,t})$ is:

$$\begin{aligned}[\boldsymbol{h}_{x_{i,0}}, \cdots, \boldsymbol{h}_{x_{i,|\boldsymbol{X}_i|}}] &= \text{Encoder}(\boldsymbol{X}_i), \\ \phi_{\text{TNorm}}(x_{i,t}) &= \left\|\boldsymbol{h}_{x_{i,t}}\right\|_2, \end{aligned} \tag{9}$$

Likewise, given a target document $\boldsymbol{Y}_i$, we obtain importance score $\phi_{\text{TNorm}}(y_{i,t})$:

$$\begin{aligned}[\boldsymbol{h}_{y_{i,0}}, \cdots, \boldsymbol{h}_{y_{i,|\boldsymbol{Y}_i|}}] &= \text{Decoder}(\boldsymbol{Y}_i, \boldsymbol{H}_{\boldsymbol{X}_i}), \\ \phi_{\text{TNorm}}(y_{i,t}) &= \left\|\boldsymbol{h}_{y_{i,t}}\right\|_2, \end{aligned} \tag{10}$$

where $\boldsymbol{H}_{\boldsymbol{X}_i} = [\boldsymbol{h}_{x_{i,0}}, \cdots, \boldsymbol{h}_{x_{i,|\boldsymbol{X}_i|}}]$. We use hidden states given by the topmost point-wise feed-forward networks in the encoder or decoder to compute the TNorm, before the layer normalization (Ba et al., 2016).

**Norm of Gradients (GNorm)** TNorm is context-aware but ignores the source-target alignment information, as $\phi_{\text{TNorm}}(x_{i,t})$ in Equation 9 does not include any information from the target document $\boldsymbol{Y}_i$. To tackle this issue, we propose to use the norm of training gradients which include all input information from both sides. Important tokens should make more contributions during the training by its gradients, resulting in larger value of gradient norm (Sato et al., 2019; Park et al., 2022). We obtain the importance score $\phi_{\text{GNorm}}(x_{i,t})$:

$$\begin{aligned}\boldsymbol{g}_{x_{i,t}} &= \nabla_{\boldsymbol{e}_{x_{i,t}}} \mathcal{L}^i(\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{\theta}), \\ \phi_{\text{GNorm}}(x_{i,t}) &= \left\|\boldsymbol{g}_{x_{i,t}}\right\|_2, \end{aligned} \tag{11}$$

where $\mathcal{L}(\boldsymbol{X}_i, \boldsymbol{Y}_i, \boldsymbol{\theta})$ is the loss function with the input of $\boldsymbol{X}_i$and $\boldsymbol{Y}_i$ seeking for the optimal parameters $\boldsymbol{\theta}$. The identical process can be directly applied to $y_{i,t}$. Note that the gradient $\boldsymbol{g}_{x_{i,t}}$ or $\boldsymbol{g}_{y_{i,t}}$ in this process is not used for updating $\boldsymbol{\theta}$.

### 3.4 Training Objective

As described in Equation 5, IADA perturbs the important information in the current sentence and accordingly increases the learning difficulty. Recent works demonstrate that *hard-to-learn* examples can hurt the model performance (Swayamdipta et al., 2020; Marion et al., 2023). To combat this issue, we draw inspiration from multi-view learning (Yan et al., 2021) and consider the perturbed samples as different views of the original samples. Therefore, we design three components in our training objective, including the *original loss*, the *perturb loss*, and the *agreement loss*:

$$\begin{aligned}\mathcal{L}^i = \; & \overbrace{\mathcal{L}^i_{\text{NLL}}(P(\boldsymbol{Y}_i|\boldsymbol{X}_i))}^{\text{original loss, see Equation 13}} + \overbrace{\mathcal{L}^i_{\text{NLL}}(P(\tilde{\boldsymbol{Y}}_i|\tilde{\boldsymbol{X}}_i))}^{\text{perturb loss}} \\ & + \underbrace{\mathcal{L}^i_{\text{JS}}(P(\boldsymbol{Y}_i|\boldsymbol{X}_i), P(\tilde{\boldsymbol{Y}}_i|\tilde{\boldsymbol{X}}_i))}_{\text{agreement loss, see Equation 14}}\end{aligned} \tag{12}$$

As defined in Equation 2, the conventional training objective of the DocNMT models for a document pair $\{\boldsymbol{X}_i, \boldsymbol{Y}_i\}$, namely the *original loss*, can be defined as:

$$\begin{aligned}&\mathcal{L}^i_{\text{NLL}}(P(\boldsymbol{Y}_i|\boldsymbol{X}_i)) = \\ &\quad -\sum \log P(y_{i,t}|\boldsymbol{y}_{i,<t}, \boldsymbol{x}_i, \boldsymbol{C}_{\boldsymbol{x}_i}, \boldsymbol{C}_{\boldsymbol{y}_i}).\end{aligned} \tag{13}$$

The *perturb loss* is defined in the same way for $\{\tilde{\boldsymbol{X}}_i, \tilde{\boldsymbol{Y}}_i\}$. Furthermore, given the equivalence between the perturbed and original samples, we introduce an extra *agreement loss*, namely Jensen-Shannon divergence:

$$\begin{aligned}&\mathcal{L}^i_{\text{JS}}(P(\boldsymbol{Y}_i|\boldsymbol{X}_i), P(\tilde{\boldsymbol{Y}}_i|\tilde{\boldsymbol{X}}_i)) = \\ &\quad \frac{1}{2}[\mathcal{D}^i_{\text{KL}}(P(\boldsymbol{Y}_i|\boldsymbol{X}_i)||P(\tilde{\boldsymbol{Y}}_i|\tilde{\boldsymbol{X}}_i)) \\ &\quad +\mathcal{D}^i_{\text{KL}}(P(\tilde{\boldsymbol{Y}}_i|\tilde{\boldsymbol{X}}_i)||P(\boldsymbol{Y}_i|\boldsymbol{X}_i))],\end{aligned} \tag{14}$$

where $\mathcal{D}^i_{\text{KL}}(\cdot||\cdot)$ is the KL divergence.

## 4 Experiments

### 4.1 Baselines

We evaluate IADA against various competitive baselines from two categories, the DocNMT baselines and the data augmentation baselines.

**DocNMT baselines** Our DocNMT baselines in this work include:

- **Doc2Doc**: The Doc2Doc baseline, proposed by Tiedemann and Scherrer (2017), incorporates contextual information into the translation process by concatenating the context and current sentence as the input for the DocNMT model.
- **HAN**: Miculicich et al. (2018) propose a hierarchical attention model to capture the contextual information. The proposed hierarchical attention encodes the contextual information in the previous sentences and have the encoded information integrated into the original NMT architecture.
- **SAN**: Maruf et al. (2019) propose the SAN baseline, which utilizes sparse attention to selectively focus on relevant sentences in the document context. It then attends to key words within those sentences.
- **Hybrid**: Zheng et al. (2020) propose the Hybrid baseline, a document-level NMT framework that explicitly models the local context of each sentence while considering the global context of the entire document in both the source and target languages.
- **FlatTrans**: The FlatTrans baseline, introduced by Ma et al. (2020), offers a simple and effective unified encoder that concatenates only the source context and the source current sentence
- **GTrans**: The GTrans baseline, proposed by Bao et al. (2021), introduces the G-Transformer, which incorporates a locality assumption as an inductive bias into the Transformer architecture.
- **MultiRes**: Sun et al. (2022) evaluate the recent DocNMT approaches and propose Multi-resolutional Training that involves multiple levels of sequence lengths.
- **DocFlat**: The DocFlat baseline, presented by Wu et al. (2023) propose Flat-Batch Attention (FBA) and Neural Context Gate (NCG) into the Transformer model.

Furthermore, we also compare our approach with a number of data augmentation approaches:

- **Word Dropout (WordDrop)** Word dropout (Gal and Ghahramani, 2016; Sennrich et al., 2016a) randomly selects a subset of positions with fixed replacement probability $p$ in an input sequence and have the selected positions replaced with $\langle \text{MASK} \rangle$.
- **Word Replacement (WordRepl)**: Word replacement (Wei and Zou, 2019; Takase and Kiyono, 2021) replaces a number of input tokens with arbitrary tokens in the vocabulary.
- **BPEDropout**: Provilkov et al. (2020) propose a simple and effective subword regularization method that randomly corrupts segmentation process of BPE.
- **CipherDAug**: Kambhatla et al. (2022) propose CipherDAug that enlarges the training data based on ROT-$k$ ciphertexts.

| | Train | Valid | Test |
|---|---|---|---|
| `TED` | $204.4K/1.7K$ | $8.9K/93$ | $2.2K/23$ |
| `News` | $242.4K/6.1K$ | $2.3K/81$ | $3.2K/155$ |
| `Europarl` | $1.8M/117.9K$ | $3.8K/240$ | $5.5K/360$ |

Table 1: The number of sentences/documents of each split of the parallel corpora.

### 4.2 Experimental Setup

**Datasets** In our experiments, we evaluated the performance of our model on three English-German translation datasets: the small-scale benchmarks `TED` (Cettolo et al., 2012) and `News Commentary`, and the large-scale benchmark `Europarl` (Koehn, 2005). For each source and target sentence, we used up to three previous sentences as the context. We tokenize the datasets with the Moses (Koehn et al., 2007) and apply BPE (Sennrich et al., 2016c) with $32K$ merges. Data statistics can be found in Table 1.

**Evaluation** We evaluate the translation quality using sentence-level SacreBLEU (Papineni et al., 2002) and document-level SacreBLEU (Liu et al., 2020), denoted as $s$-BLEU and $d$-BLEU.[1] To assess the contextual awareness of DocNMT models, we employ the English-German anaphoric pronoun test set introduced by Müller et al. (2018). This test requires the model to identify the correct pronoun (*er*, *es*, or *sie*) in German among several candidate translations, and the performance is mea-

[1] SacreBLEU signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.0`.

| | TED | | | News | | | Europarl | | |
|---|---|---|---|---|---|---|---|---|---|
| | *s*-BLEU | *d*-BLEU | COMET | *s*-BLEU | *d*-BLEU | COMET | *s*-BLEU | *d*-BLEU | COMET |
| HAN (2018) | 24.6 | — | — | 25.0 | — | — | 28.6 | — | — |
| SAN (2019) | 24.4 | — | — | 24.8 | — | — | 29.7 | — | — |
| HYBRID (2020) | 25.1 | — | — | 24.9 | — | — | 30.4 | — | — |
| FLATTRANS (2020) | 24.9 | — | — | 23.6 | — | — | 30.1 | — | — |
| GTRANS (2021) | 25.1 | 27.2 | — | 25.5 | 27.1 | — | 32.4 | 34.1 | — |
| MULTIRES (2022) | 25.2 | 29.3 | — | 25.0 | 26.7 | — | 32.1 | 34.5 | — |
| DOCFLAT (2023) | 25.4 | — | **31.0** | 25.4 | — | 21.2 | 32.2 | — | 59.9 |
| WORDDROP (2016a) | 24.5 | 28.1 | 26.6 | 24.5 | 26.7 | 16.9 | 31.6 | 33.7 | 59.0 |
| WORDREPL (2019) | 24.6 | 28.5 | 27.7 | 24.9 | 26.9 | 18.0 | 31.9 | 33.8 | 58.9 |
| BPEDROPOUT (2020) | 25.1 | 28.9 | 28.8 | 25.6 | 27.4 | 20.3 | 32.2 | 34.0 | 59.9 |
| CIPHERDAUG (2022) | 24.2 | 28.0 | 19.7 | 24.4 | 26.7 | 14.4 | 31.4 | 33.2 | 58.5 |
| *Importance-Aware Augmented (Ours)* | | | | | | | | | |
| DOC2DOC (doc baseline) | 24.3 | 27.4 | 23.5 | 24.4 | 26.4 | 12.7 | 31.2 | 33.1 | 58.4 |
| + IADA$_{\text{DROP}}$ + TNORM | 25.6 | 29.3 | 28.7 | 26.2 | 28.3 | 20.1 | 32.7 | 34.9 | 60.3 |
| + GNORM | 26.1 | 29.6 | 29.8 | 26.3 | 28.6 | 20.7 | 32.8 | 35.0 | 60.3 |
| + IADA$_{\text{REPL}}$ + TNORM | 26.1 | **29.7** | 29.7 | 26.3 | 28.5 | 20.8 | 32.8 | 34.8 | 60.3 |
| + GNORM | **26.2** | 29.6 | 29.8 | **26.4** | **28.7** | **22.1** | **33.0** | **35.1** | **60.4** |
| *Fine-tuning from pre-trained models for comparison* | | | | | | | | | |
| FLATTRANS + BERT | 26.6 | — | — | 24.5 | — | — | 32.0 | — | — |
| GTRANS + BERT | 26.8 | — | — | 26.1 | — | — | 32.4 | — | — |
| GTRANS + MBART | 28.0 | 30.0 | — | 30.3 | 31.7 | — | 32.7 | 34.3 | — |

Table 2: Main results on English-German document-level machine translation. All the results given by IADA$_{\text{DROP}}$ and IADA$_{\text{REPL}}$ significantly outperform DOC2DOC at the significance level $p = 0.05$ based on Koehn (2004). Best results are highlighted in **bold**.

sured by *Accuracy*.

**Inference** We translate test examples in their original order, beginning with the first sentence independent of context. Previous translations serve as the context for the current translation.

**Hyperparameters** All the approaches in this works, including IADA and baselines, are trained from scratch with the identical hyperparameters. The model is randomly initialized and optimized with Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and the learning rate $\alpha = 5 \times 10^{-4}$. The model is trained with the batch size of $32K$ tokens for both datasets and the dropout rate $p = 0.3$. The batch size of $32K$ tokens is achieved by using the batch size of 4096 tokens and updating the model for every 8 batches. The learning rate schedule is the same as described in Vaswani et al. (2017) with $4K$ warmup steps. We use early stopping on validation loss. For our IADA approach, we set the initial replacement probabilities for both the context and the current sentence to be $p_{\text{ctx}} = p_{\text{cur}} = 0.1$. We set the $\alpha$ in Equation 6 to $\alpha = 0.1$.

**Computational Infrastructure** The model architecture for all the approaches in this work is Transformer-base (Vaswani et al., 2017), having about $64M$ parameters. We run experiments with two A100 GPUs. Each experiment for IADA on TED commonly take less than 5 hours. The computational cost of IADA on News and Europarl is proportional to that of TED with regard to the size of training corpus.

### 4.3 Main Result

We present the main results in Table 2.

**Comparison with other approaches** Our IADA$_{\text{DROP}}$ and IADA$_{\text{REPL}}$ models surpass other DOCNMT models in performance without requiring additional neural modules or incurring computational overhead. Moreover, IADA models also outperform other competitive DA approaches on both *s*-BLEU and *d*-BLEU. They exhibit substantial performance gains on all three benchmarks, demonstrating their effectiveness in training DOCNMT models for both low-resource and high-resource settings. In contrast, other DA approaches only exhibit marginal improvements on the large benchmark Europarl.

**IADA$_{\text{DROP}}$ vs. IADA$_{\text{REPL}}$** Both IADA$_{\text{DROP}}$ and IADA$_{\text{REPL}}$ consistently outperform the unaugmented DOC2DOC baseline, WORDDROP, and WORDREPL, with statistical significance, demonstrating the effectiveness of our method. Interest-

| | Cont. | Curr. | $s$-B. | $d$-B. | $s$-C. |
|---|---|---|---|---|---|
| WordRepl | — | — | 24.6 | 28.5 | 27.7 |
| Doc2Doc | — | — | 24.3 | 27.4 | 23.5 |
| + IADA$_{\text{Repl}}$ + TNorm | ↓ | ↑ | **26.1** | **29.7** | **29.7** |
| | ↑ | ↓ | 25.7 | 28.9 | 29.3 |
| | ↓ | ↓ | 25.6 | 28.4 | 29.5 |
| | ↑ | ↑ | 25.5 | 28.3 | 28.9 |
| + IADA$_{\text{Repl}}$ + GNorm | ↓ | ↑ | **26.2** | **29.6** | **29.8** |
| | ↑ | ↓ | 25.8 | 28.5 | 29.4 |
| | ↓ | ↓ | 25.9 | 28.8 | 29.2 |
| | ↑ | ↑ | 25.7 | 28.6 | 29.3 |

Table 3: Ablation study for the perturbation strategy in Equation 5 given by IADA$_{\text{Repl}}$ on TED. Best results are highlighted in **bold**. ↑ indicates $s+\psi(x_{i,t})$ or $s+\psi(y_{i,t})$. ↓ indicates $s-\psi(x_{i,t})$ or $s-\psi(y_{i,t})$.

| | $s$-BLEU | $d$-BLEU | COMET |
|---|---|---|---|
| WordRepl | 24.6 | 28.5 | 27.7 |
| *Normalized* | | | |
| Doc2Doc | 24.3 | 27.4 | 23.5 |
| + IADA$_{\text{Repl}}$ + TNorm | 26.1 | **29.7** | 29.7 |
| + IADA$_{\text{Repl}}$ + GNorm | **26.2** | 29.6 | 29.8 |
| + IADA$_{\text{Repl}}$ + Random | 24.6 | 28.4 | 25.4 |
| *Not normalized* | | | |
| + IADA$_{\text{Repl}}$ + TNorm | 24.5 | 27.8 | 25.0 |
| + IADA$_{\text{Repl}}$ + GNorm | 24.7 | 27.9 | 25.2 |

Table 4: Ablation study for token importance measures and token importance normalization given by IADA$_{\text{Repl}}$ on TED. Best results are highlighted in **bold**.

ingly, we observe that WordRepl-based methods (IADA$_{\text{Repl}}$ and WordRepl) slightly outperform the WordDrop-based methods (IADA$_{\text{Drop}}$ and WordDrop). We hypothesize that WordRepl-based methods generate more diverse synthetic data by replacing selected tokens with distinct random tokens, compared with replaceing selected tokens with ⟨mask⟩. Lastly, we also observe that GNorm outperforms TNorm, confirming our hypothesis in Section 3.3.

## 4.4 Ablation Study

In this section, we conduct ablation studies to show the effectiveness of IADA components based on IADA$_{\text{Repl}}$ on the TED benchmark.

**Perturbation Strategy** Our proposed perturbation strategy's effectiveness is demonstrated by enumerating all possible strategies for token importance measures in Table 3. For instance, ↑ for the context and ↓ for the current sentence in Table 3 indicate a tendency to perturb *important* information in the context while perturbing *less important* information in the current sentence. Results consistently indicate that all other perturbation strategies are suboptimal compared to our strategy. This success is attributed to the design of IADA, which encourages DocNMT models to leverage contextual information.

| | $s$-BLEU | $d$-BLEU | COMET |
|---|---|---|---|
| WordRepl | 24.6 | 28.5 | 27.7 |
| Doc2Doc | 24.3 | 27.4 | 23.5 |
| + IADA$_{\text{Repl}}$ + TNorm | **26.1** | **29.7** | **29.7** |
| - anchor loss | 25.2 | 28.8 | 28.5 |
| - perturb loss | 25.5 | 28.4 | 28.3 |
| - agreement loss | 25.4 | 28.5 | 28.5 |
| + IADA$_{\text{Repl}}$ + GNorm | **26.2** | **29.6** | **29.8** |
| - anchor loss | 25.3 | 28.7 | 29.0 |
| - perturb loss | 25.4 | 28.3 | 28.5 |
| - agreement loss | 25.6 | 28.5 | 28.3 |

Table 5: Ablation study for the loss terms in Equation 12 given by IADA$_{\text{Repl}}$ on TED. "-" indicates removing the loss term. Best results are highlighted in **bold**.

**Token Importance Measures** To demonstrate the effectiveness of our proposed importance measures, we replace $\psi(\cdot)$ in Equation 5 with a random score $r \sim \mathcal{N}(0, \alpha^2)$ according to Equation 6. This method is referred to as Random in Table 4. We observe that IADA$_{\text{Repl}}$ with Random achieves performance similar to WordRepl, suggesting that the importance measures can more effectively guide the generation of high-quality synthetic data compared to purely random approaches.

**Importance Normalization** We examine the impact of importance normalization (Equation 6) in Table 4. Without this normalization, both IADA$_{\text{Repl}}$ with TNorm and IADA$_{\text{Repl}}$ with GNorm experience notable performance declines and slightly underperform the WordRepl baseline. These findings emphasize the crucial role of controlling the spread of $\phi(x_{i,t})$ in IADA.

**Training Objective** We analyze the effectiveness of each loss term of Equation 12 and present our findings in Table 5. Our results demonstrate that each loss term plays a significant role in improving the model performance. Notably, when we remove the *perturb loss*, we observe a greater decrease in $d$-BLEU, indicating that our IADA design effectively encourages the model to utilize the context to enhance document-level translation quality.

| | Acc. | *er* | *es* | *sie* |
|---|---|---|---|---|
| WORDREPL | 68.0 | 56.6 | **92.0** | 55.5 |
| DOC2DOC | 63.5 | 51.2 | 89.6 | 49.9 |
| + IADA$_{\text{REPL}}$ + TNORM | 71.2 | 58.3 | 90.8 | 64.3 |
| + GNORM | **73.8** | **63.9** | 89.4 | **67.8** |

Table 6: Accuracy (in %) on the contrastive test set given by IADA$_{\text{REPL}}$ trained on TED. Best results are highlighted in **bold**.

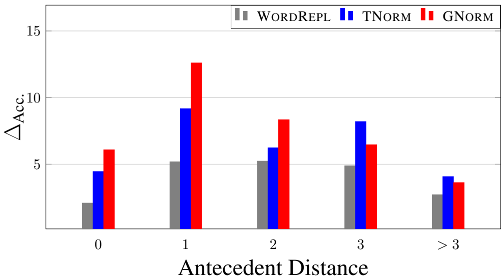


Figure 3: Accuracy gap (in %; $\Delta_{\text{Acc.}}$) given by WORDREPL and IADA$_{\text{REPL}}$ with different token importance measures on TED against DOC2DOC with regard to the antecedent distance (in sentences).

| | $s$-BLEU | $d$-BLEU | COMET |
|---|---|---|---|
| DOCBT | 25.0 | 28.8 | 28.9 |
| DOCFT | 25.1 | 28.9 | 29.0 |
| DOC2DOC | 24.3 | 27.4 | 23.5 |
| + IADA$_{\text{REPL}}$ + TNORM | 26.1 | 29.6 | 29.7 |
| + DOCBT | 26.8 | 30.2 | 30.5 |
| + DOCFT | 26.9 | 30.4 | 31.0 |
| + IADA$_{\text{REPL}}$ + GNORM | 26.2 | 29.6 | 29.8 |
| + DOCBT | 26.6 | 30.1 | 30.7 |
| + DOCFT | 26.9 | 30.6 | 31.1 |

Table 7: Compatibility with DOCBT and DOCFT of IADA$_{\text{REPL}}$ on TED. Best results are highlighted in **bold**.

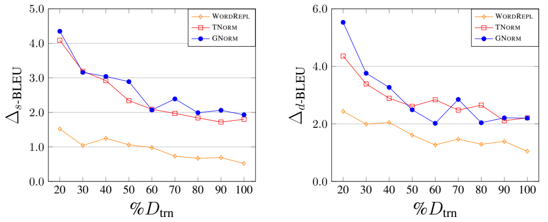


Figure 4: The performance gap ($\Delta_{\{\cdot\}}$) given by IADA$_{\text{REPL}}$ and WORDREPL against DOC2DOC with regard to the percentage of training data ($\%D_{\text{trn}}$) of TED.

## 5 Analysis

We analyze IADA from various aspects in this section, including contextual awareness, robustness, compatibility with DOCBT/DOCFT, simulated low-resource scenario, and linguistic analysis.

**Contextual Awareness** In our analysis, we evaluate the contextual awareness of DOCNMT models using a contrastive test set. We focus on the accuracy of different *anaphoric pronoun types* (Table 6) and *antecedent distance* (Figure 3). The choice of anaphoric pronoun types, such as feminine *sie*, neutral *er*, and masculine *es*, depends on the context in English-German translation. Results in Table 6 demonstrate that IADA$_{\text{REPL}}$ achieve higher overall accuracy compared with DOC2DOC and WORDREPL. These improvements mainly come from the minor classes, feminine *sie* and neutral *er*, indicating that IADA effectively overcomes the training bias towards the major class *es*. Regarding the antecedent distance shown in Figure 3, both IADA$_{\text{REPL}}$ with TNORM and IADA$_{\text{REPL}}$ with GNORM consistently outperform WORDREPL across all distances.

**Compatibility with DOCBT/DOCFT** We investigate the compatibility of IADA with back-translation (DOCBT) and forward-translation (DOCFT). We start from doubling the original training corpus using DOCBT or DOCFT and then augmenting it with IADA. The results in Table 7 demonstrate that combining IADA$_{\text{REPL}}$ variants with DOCBT and DOCFT yields further improvements. The hybrid systems outperform both individual systems, indicating the successful integration of IADA with DOCBT and DOCFT.

**Simulated Low-Resource Scenario** We also examine the usefulness of IADA in low-resource training scenarios. We vary the size of the training data ($D_{\text{trn}}$) for TED from 20% (around 40K) to 100% (around 200K). The performance gap ($\Delta_{\{\cdot\}}$) compared to the DOC2DOC model is shown in Figure 4 for all three metrics. Overall, IADA$_{\text{REPL}}$ variants with TNORM, and GNORM outperform WORDREPL across different data scales. In particular, When using only 20% of the TED training data, IADA$_{\text{REPL}}$ with GNORM achieves approximately +4.5 and +5.5 improvements in $s$-BLEU and $d$-BLEU respectively compared to DOC2DOC, while WORDREPL provides only a +1.5 and +2.5 improvements for $s$-BLEU and $d$-BLEU. These results highlight the effectiveness of IADA in various low-resource data scenarios.

**Robustness against Noisy Context** In our experiment, we test the effectiveness of IADA in mitigating negative impacts of irrelevant and dis-

| | *s*-BLEU | | | *d*-BLEU | | | COMET | | | Accuracy | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gold | Noisy | Δ ↓ | Gold | Noisy | Δ ↓ | Gold | Noisy | Δ ↓ | Gold | Noisy | Δ ↓ |
| DOC2DOC | 24.3 | 23.5 | 0.8 | 27.4 | 26.3 | 1.1 | 23.5 | 22.0 | 1.5 | 63.5 | 46.8 | 16.7 |
| WORDREPL | 24.6 | 24.0 | 0.6 | 28.5 | 27.8 | 0.7 | 27.7 | 26.2 | 1.5 | 68.0 | 53.4 | 14.6 |
| IADA$_{\text{REPL}}$ | | | | | | | | | | | | |
| +TNORM | 26.1 | 25.7 | **0.4** | **29.7** | **29.4** | 0.3 | 29.7 | 28.9 | 0.8 | 71.2 | 63.1 | 8.1 |
| +GNORM | **26.2** | **25.8** | **0.4** | 29.6 | **29.4** | **0.2** | **29.8** | **29.3** | **0.5** | **73.8** | **66.0** | **7.8** |

Table 8: Performance gap (Δ) given by the selected methods trained with the gold context against the noisy context on TED. Best results are highlighted in **bold**. ↓ indicates lower is better.

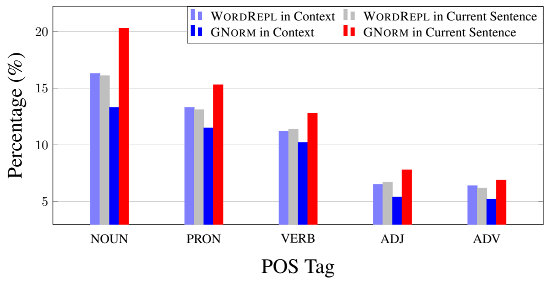


Figure 5: The percentage (%) of the POS tags of the perturbed tokens on TED given by WORDREPL and IADA$_{\text{REPL}}$ with GNORM.

ruptive context. We randomly replace two out of three sentences in the gold context of training instances with sentences from other documents. Results on TED (Table 8) shows that IADA$_{\text{REPL}}$ variants have smaller performance declines compared to WORDREPL and DOC2DOC. Notably, even with noisy context, IADA$_{\text{REPL}}$ with GNORM outperforms DOC2DOC with gold context across all metrics. Our preliminary study shows that a vanilla sentence-level Transformer-base model trained on TED achieves approximately 45% accuracy. The decline in accuracy for DOC2DOC suggests its susceptibility to noisy context. Overall, IADA successfully trains DOCNMT models to focus on relevant context and enhances their robustness with low-quality input information.

**Linguistic Analysis on Perturbed Tokens** We analyze perturbed tokens from WORDREPL and IADA$_{\text{REPL}}$ with GNORM using linguistic analysis, focusing on five significant Part-Of-Speech (POS) tags. The results (Figure 5) reveal that compared to WORDREPL, IADA$_{\text{REPL}}$ with GNORM consistently selects more tokens with major POS tags in the current sentence, while IADA$_{\text{REPL}}$ perturbs fewer tokens with major POS tags in the context. These findings confirm that IADA prioritizes perturbing important tokens in the current sentence and the less important ones in the context.

## 6 Conclusion

In this paper, we present IADA, a new method for generating high-quality syntactic data for DOCNMT. By leveraging *token importance*, IADA augments existing training data by perturbing important tokens in the current sentences while keeping those less important ones unchanged. This encourages DOCNMT models to effectively utilize contextual information. We propose TNORM and GNORM to measure token importance. We also introduce the agreement loss to prevent the training samples from being overly hard to learn after perturbation. Results demonstrate that IADA outperforms competitive DOCNMT approaches as well as several data augmentation methods. Our analysis reveals that IADA enhances DOCNMT models' contextual awareness, robustness, and is compatible with DOCBT and DOCFT techniques. IADA also shows significant benefits in low-resourced settings. Linguistic analysis validates the effectiveness of IADA in identifying important tokens. Overall, our findings highlight the efficacy of IADA in improving syntactic data generation for DOCNMT.

## 7 Limitations

Comparing with standard optimization techniques, our proposed IADA with the TNORM and GNORM requires additional forward and backward computation. For each training step, IADA with TNORM requires one additional forward pass, and IADA with GNORM requires one additional forward and backward pass. Note that IADA is only applied to the training stage and has no impact on the DOCNMT inference.

## References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Guangsheng Bao, 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. Association for Computational Linguistics.

Rachel Bawden, 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. Association for Computational Linguistics.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Verna Dankers, Elia Bruni, and Dieuwke Hupkes. 2022. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, 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. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware 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 6467–6478, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1019–1027.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2020. Revisiting self-training for neural sequence generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Jingjing Huo, 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. Association for Computational Linguistics.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Nishant Kambhatla, Logan Born, and Anoop Sarkar. 2022. CipherDAug: Ciphertext based data augmentation for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 201–218, Dublin, Ireland. Association for Computational Linguistics.

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

Philipp Koehn. 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. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion*

*Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo. 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. Association for Computational Linguistics.

Yikun Lei, 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. International Committee on Computational Linguistics.

Pengfei Li, Liangyou Li, Meng Zhang, Minghao Wu, and Qun Liu. 2022. Universal conditional masked language pre-training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6379–6391, Dublin, Ireland. Association for Computational Linguistics.

Qi Liu, Matt Kusner, and Phil Blunsom. 2021. Counterfactual data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 187–197, Online. Association for Computational Linguistics.

Yinhan Liu, 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.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022. Divide and rule: Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4557–4572, Dublin, Ireland. Association for Computational Linguistics.

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181*.

Shuming Ma, 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. Association for Computational Linguistics.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *CoRR*, abs/2309.04564.

Sameen Maruf 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. Association for Computational Linguistics.

Sameen Maruf, 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. Association for Computational Linguistics.

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.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

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

Jungsoo Park, Gyuwan Kim, and Jaewoo Kang. 2022. Consistency training with virtual adversarial discrete perturbation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5646–5656, Seattle, United States. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Motoki Sato, Jun Suzuki, and Shun Kiyono. 2019. Effective adversarial regularization for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 204–210, Florence, Italy. Association for Computational Linguistics.

Adriaan M. J. Schakel and Benjamin J. Wilson. 2015. Measuring word significance using distributed representations of words. *CoRR*, abs/1508.02297.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. 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.

Connor Shorten, Taghi M. Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *J. Big Data*, 8(1):101.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Sho Takase and Shun Kiyono. 2021. Rethinking perturbations in encoder-decoders for fast training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5767–5780, Online. Association for Computational Linguistics.

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.

Sami Ul Haq, Sadaf Abdul Rauf, Arsalan Shaukat, and Abdullah Saeed. 2020. Document level NMT of low-resource languages with backtranslation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 442–446, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. 2022. PromDA: Prompt-based data augmentation for low-resource NLU tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4242–4255, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Benjamin J. Wilson and Adriaan M. J. Schakel. 2015. Controlled experiments for word embeddings. *CoRR*, abs/1510.02675.

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.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document flattening: Beyond concatenating context for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 448–462, Dubrovnik, Croatia. Association for Computational Linguistics.

Minghao Wu, Yitong Li, Meng Zhang, Liangyou Li, Gholamreza Haffari, and Qun Liu. 2021. Uncertainty-aware balancing for multilingual and multi-domain neural machine translation training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7291–7305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. 2021. Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Austin, Texas. Association for Computational Linguistics.

Long Zhang, Tong Zhang, Haibo Zhang, Baosong Yang, Wei Ye, and Shikun Zhang. 2021. Multi-hop transformer for document-level machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3953–3963, Online. Association for Computational Linguistics.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3983–3989. ijcai.org.