Quality Estimation based feedback training for improving pronoun translation

Harshit Dhankhar
Indian Institute of Technology Patna
harshit_2101mc20@iitp.ac.in

Baban Gain
Indian Institute of Technology Patna
gainbaban@gmail.com

Asif Ekbal
Indian Institute of Technology Jodhpur
asif@iitj.ac.in

Y.M. Tripathi
Indian Institute of Technology Patna
yogesh@iitp.ac.in

Abstract

Pronoun translation is a longstanding challenge in neural machine translation (NMT), often requiring inter-sentential context to ensure linguistic accuracy. To address this, we introduce ProNMT, a novel framework designed to enhance pronoun and overall translation quality in context-aware machine translation systems. ProNMT leverages Quality Estimation (QE) models and a unique Pronoun Generation Likelihood-Based Feedback mechanism to iteratively fine-tune pre-trained NMT models without relying on extensive human annotations. The framework combines QE scores with pronoun-specific rewards to guide training, ensuring improved handling of linguistic nuances. Extensive experiments demonstrate significant gains in pronoun translation accuracy and general translation quality across multiple metrics. ProNMT offers an efficient, scalable, and context-aware approach to improving NMT systems, particularly in translating context-dependent elements like pronouns.

1 Introduction

Document translation is a critical application of machine translation (MT), facilitating cross-lingual transfer of knowledge. In an increasingly interconnected world, multilingual communication is essential to ensure equitable access to information and services. Despite advances in neural machine translation (NMT) models and the rise of large language models (LLMs), document translation remains a challenging and relevant area of research (Sun et al., 2022; Wang et al., 2023). Traditional MT systems often process sentences independently, which can lead to inconsistencies in terminology and style across a document.

Document translation addresses these limitations by leveraging contextual information across sentences, paragraphs, or entire documents to produce more coherent and accurate translations. For instance, incorporating document-level context improves the handling of anaphora (Voita et al., 2018), lexical disambiguation, and stylistic consistency.

Various techniques have been employed in literature to fine-tune MT models, with the most prominent ones being: a) Supervised fine-tuning (SFT), which uses labeled data to fine-tune the model in a supervised fashion, and b) Reinforcement Learning from human feedback (RLHF) which is based on optimizing a reward function based on expert judge rankings (human preferences) and using it with Proximal Policy Optimization (PPO) to fine-tune the model. One of the major drawbacks of both methods is the explicit dependence on human experts to either label huge datasets or rank candidate translations. Furthermore, training PPO involves tuning a large set of hyperparameters and loading multiple models (reference, critic, and reward), which comes at the expense of computational power and expansive memory resources. Although less stable and faster than SFT, RLHF using PPO has shown superior performance in aligning models (Ramamurthy et al., 2023).

Various attempts have been made to integrate feedback to improve the quality of translations in the field of neural machine translation (NMT) as well. Although a few works employ real but limited human feedback (Kreutzer et al., 2018a,Kreutzer et al., 2018b), others focus on using similarity scores between candidates and reference translation as a simulated human feedback. Quality Estimation (QE) models have recently been proven to be an adept proxy for real human feedback-based reward models (He et al., 2024). These QE models, facilitated by the advent of more human evaluation data and better language models (Rei et al., 2020), provide a numerical score to indicate the quality of candidate translation. Our proposed framework is based on exploiting these QE model evaluations to assist the feedback training process iteratively, bypassing the requirement to perform human evaluations since it is very costly in most cases.

Despite substantial progress in various areas related to neural machine translation (NMT), the task of translating pronouns has always been inherently difficult for MT models due to dependence on inter-sentential context for their translation. For example, see the case presented in Fig 1. In both languages, a pronoun in the second sentence refers to advertising. Hence, when the second sentence is translated from English to German, the translation of the pronoun *it* is ambiguous without the previous sentence. This calls for the need for a framework that can help fine-tune MT models to give better overall and pronoun translation scores.

Existing methods for document translation often incorporate techniques such as architectural-level modifications to include document-level context, concatenation of context sentences, and multitasking approaches. These methods aim to enhance the translation quality by leveraging the additional contextual information available in documents. However, most models are traditionally trained using reference translations, without explicitly focusing on key areas where document-level machine translation (MT) systems excel—such as improving the translation of pronouns, which are often challenging in standard sentence-level MT systems.

In this work, we introduce **ProNMT**, a novel framework designed to exploit the capabilities of context-aware MT models to improve translation quality, with a particular emphasis on pronoun translation. ProNMT is built upon two core ideas:

1. **Quality Estimation (QE)**: This component evaluates the translation quality of pronouns in the output by estimating how well the generated pronouns align with contextual and linguistic expectations.

2. **Pronoun Generation Likelihood-Based Feedback**: This is a unique training mechanism where feedback is provided based on the likelihood of generating correct pronouns during translation.

We define **"pronoun generation likelihood"** as the probability assigned by the model to a pronoun token, given the source sentence and the previously generated tokens. This metric serves as a proxy for assessing the quality of pronoun translation, as detailed in Section 3.2. By incorporating this likelihood into our feedback loop, we aim to guide the model toward improved pronoun handling while simultaneously enhancing the overall translation quality.

This paper makes the following key contributions:

1. **Pronoun-Focused Feedback Mechanism**: We propose a feedback mechanism specifically tailored to enhance pronoun translation in context-aware MT systems. This mechanism integrates pronoun generation likelihood as a measurable and actionable metric during training.

2. **Context-Aware Quality Enhancement**: We exploit document-level context to improve both pronoun translation and the general quality of translations, demonstrating how a targeted approach to pronoun handling can benefit overall translation performance.

3. **Novel Framework for Fine-Tuning Pre-Trained MT Models**: ProNMT provides a practical framework for fine-tuning pre-trained MT models, leveraging quality estimation and feedback-driven training processes to address long-standing challenges in document-level MT.

4. **Evaluation and Results**: We empirically validate our approach through extensive experiments, highlighting significant improvements in pronoun translation and overall translation quality compared to existing methods.

By addressing the specific challenges of pronoun translation and leveraging document-level context, ProNMT sets a new benchmark for improving the quality of translations in pre-trained machine translation models.

EN This advertising cheapens all women.
It cheapens every one of us and our daughters.

DE Eine solche Art von Werbung erniedrigt alle Frauen.
Sie erniedrigt uns alle und unsere Töchter.

Figure 1: Example illustrating the inter-sentential dependence for pronoun translation. Pronouns of interest are *in italics*, and the antecedents they refer to are underlined. Data taken from Europarl EN <-> DE dataset.

2 Related Works

Incorporating context is generally better than context-agnostic models (Sim Smith, 2017). The primary methods to incorporate contexts often use either concatenation (Tiedemann and Scherrer, 2017; Junczys-Dowmunt, 2019) or multi-encoder-based approaches. Multi-encoder architectures, while helps to achieve better results, similar results could be obtained by passing random context instead of actual context in the additional encoder (Li et al., 2020). Appicharla et al. (2024) explored multi-task learning (MTL) in context-aware NMT by explicitly modeling context encoding to enhance sensitivity to context choice. Experiments on German-English language pairs showed that the MTL approach outperformed concatenation-based and multi-encoder DocNMT models in low-resource settings. However, they observed that MTL models struggled to generate the source from the context, suggesting that available document-level parallel corpora may not be sufficiently context-aware. (Wang et al., 2020) used previous three sentences during pre-training of Cross-lingual Language model Pre-training. Translating pronouns accurately in Neural Machine Translation (NMT) systems remains a significant challenge, primarily due to the necessity of utilizing inter-sentential context. Similarly, (Appicharla et al., 2023) investigated the impact of different context settings on pronoun translation accuracy. They trained multi-encoder models using previous sentences, random sentences, and a mix of both as context, evaluating their performance on the ContraPro test set. Their models performed well even with random context, indicating that the models were somewhat agnostic to the specific context provided. Voita et al. (2018) observed that using document-level context helps in better pronoun translation. While the previous works observed the effectiveness of context in translating pronouns, the effect of pronoun translation as one of the objective is yet to be explored.

Our work differs from these approaches by introducing **ProNMT**, a framework that leverages Quality Estimation (QE) models and a Pronoun Generation Likelihood-Based Feedback mechanism to iteratively fine-tune pre-trained NMT models. Unlike previous methods that rely on modeling context through auxiliary tasks or synthetic data, ProNMT bypasses the need for extensive human annotations and explicitly focuses on improving pronoun translation within context-aware systems. By integrating QE scores with pronoun-specific rewards, our method effectively guides the training process to enhance pronoun handling and overall translation quality, offering a scalable solution to longstanding challenges in document-level translation.

3 Methodology

3.1 Framework

Given a pre-trained MT model $M_0(x, \theta_0)$ with initial parameters θ_0, which generates an output y based on multinomial sampling with underlying distribution $p_{M_0}(y|x, \theta_0)$, our aim is to guide the model to generate better translations with a focus on pronouns using a QE-based reward function $r(x, y)$. Note that the QE-based reward function does a reference-free estimation of the translation quality. We define the optimization objective as:

$$\max_{\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim p_M(y|x;\theta)} r(x, y). \tag{1}$$

In each iteration i, we choose a batch of sentence pairs $\{X_i, Y_i\}$ of size B. For each sample input $x_p \epsilon X_i$, we then generate k candidate translations $y_1^p, y_2^p, ...y_k^p$. Then, we extract the pronoun of interest from each candidate $y_j^p\ \forall j$ and calculate the reward as defined in (2). This helps our framework to choose the best proxy for human feedback (best candidate translation, say y_j^p) and update the model parameters using iterative supervised fine-tuning (SFT). This iterative training process is presented in Algorithm 1.

3.2 Reward function

The reward function used for training is a linear combination of pronoun-based and translation-based reward metrics. For a given pair of sentences (x_p, y_p^t) containing the target pronoun token y_{pi}^t and a generated candidate translation y_k^p, we assess overall translation quality using the COMET model "wmt21-comet-qe-da" that employs a reference-free evaluation approach and is built on the XLM-R architecture (Rei et al., 2020). This gives us $R_{translation}$, a normalized score between -1 and 1, where 1 means perfect translation. Next, to assess the pronoun translation reward R_{PGL}, we identify the pronoun token y_{kj}^p in candidate translation y_k^p and its "Pronoun Generation Likelihood (PGL)" defined as: $P(y_{kj}^p \mid \mathbf{x_p}, \mathbf{y_k}_{1:j-1}^p; \theta)$. If the pronoun token matches with that in reference translation, then we set the pronoun reward to PGL itself. If it does not, then it is set to -PGL. Lastly, if no pronoun token in present in the candidate in the first place, R_{PGL} is set to 0. We are now in a position to define the overall reward $r(x, y_j)$ in Equation 2.

$$r(x, y_j) = \beta \cdot R_{PGL} + \alpha R_{translation} \tag{2}$$

where

$$R_{PGL} = \begin{cases} PGL & \text{if } y^p_{kj} = y^t_{pi}, \\ -PGL & \text{if } y^p_{kj}! = y^t_{pi}, \\ 0 \text{ otherwise} & \end{cases}$$

Algorithm 1 ProNMT

Require: Training set $\mathcal{X}$, reward function $r(x, y)$, initial model $M_0 = P(y|x; \theta_0)$, batch size B, temperature T, the number of candidates k

for iteration i in $0, 1, \ldots, N-1$ **do**
  $D_i \leftarrow$ SampleBatch$(\mathcal{X}, b)$
  $\mathcal{B} \leftarrow \emptyset$
  for each $x \in D_i$ **do**
    $y_1, \ldots, y_k \sim P_T(y|x; \theta_i)$
    $y^* \leftarrow \arg\max_{y_j \in \{y_1, \ldots, y_k\}} r(x, y_j)$
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{(x, y^*)\}$
  end for
  Fine-tune θ_i on $\mathcal{B}$ to obtain $M_{i+1} = P(y|x; \theta_{i+1})$
end for

4 Experiments

4.1 Data for training

While a contrastive test-suite to assess pronoun translation quality called Contrapro (Müller et al., 2018) is available, we found it to contain relatively shorter and easy-to-translate source sentences which makes it trivial for fine-tuning a pre-trained MT model (see appendix section B). This motivated us to design our own Europarl-based filtered dataset. We start preparing our training data by adopting Europarl our base EN <-> DE sentence corpus. It contains 1920209 sentence pairs of English and German languages. To make the dataset suited for pronoun translation we adopt the following filtering process: for each pair of sentences (s, t) in English and German, extract iff

- s contains the English pronoun it, and t contains a German pronoun that is third person singular (er, sie or es), as indicated by their part-of-speech tags.
- Those pronouns are aligned to each other.

Note that we only consider the pronoun "it" and its German translations "er" (Masculine), "sie" (Feminine) or "es" (Neutral) due to the crucial dependence on source or target side context for its translation quality. This corpus filtering process is also crucial to reduce noise in QE-based feedback training. If all pronouns are considered in the filtered one-pronoun sentences, we observed the training to be noisy due to the possible ambiguity in pronoun translation, i.e. more than one candidate pronoun translations may be valid (see appendix section A.1 for details). Further, checking alignment of pronouns is also important as "it" may correspond to different german pronouns like "ist","das" or "dies" as well (Müller et al., 2018). This filtering process reduces the dataset to 117834 sentences containing "es", 17447 sentences containing "er" and 39439 sentences containing "sie". To tackle class imbalance in the filtered dataset, we sample 15000 sentences from each class to create our final dataset used for training. The final train set contains 42750 sentences, test set contains 1500 sentences and validation set contains 750 instances. For our document-level experiment, we preprocess the source sentence x_i in the following format: "<context> x_{i-1} <\context> x_i". Note that the same seed was set for shuffling data instances while creating train,test and val sets, for both with and without context experiments. This was done to maintain uniformity in performing assessments.

4.2 Training details

We begin the training process by considering a base pre-trained model. For our experiments, we chose the distilled 600M parameter variant of NLLB-200. In this model, we deploy an iterative Supervised Fine Tuning (SFT) trainer using the trl package from transformers library. This helps us to fine tune the model parameters on the reward-chosen candidates iteratively for each mini-batch. We choose the batch size $B = 4$ for our experiments. For each input sentence x_p in a mini-batch, we generate $k = 10$ candidate translation $(y^p_1, y^p_2 ... y^p_{10})$ through multinomial sampling. The "most-suited" candidate y^c_i is then greedily chosen for each sample in the mini-batch and fed into the Iterative-SFT Trainer to update the model parameters. We run the trainer for a maximum of 700 iterations with a learning rate of 6e-5 and gradient accumulation steps = 16. For assessing the model in a document-level setting, we first fine-tuned the model on a subset of our filtered dataset using Huggingface's trainer class. For fine-tuning, we train the model for 3 epochs with the hyperparameters mentioned in Table 1. This is done so that model learns to use the context but not output it in translation during generation. We perform SFT training for without-context model on same hyperparameters as well, although not required, it helps to normalise testing scores across with and without-context trained models.

Hyperparameter	Value
Learning Rate	8×10^{-3}
Per Device Train Batch Size	8
Gradient Accumulation Steps	16
Eval Accumulation Steps	16
Per Device Eval Batch Size	8
Weight Decay	0.01
Number of Train Epochs	10

Table 1: Hyperparameters for fine-tuning process

4.3 Evaluation

We chose to assess the model trained through ProNMT during training and testing. For the testing evaluation, we assess the model checkpoint with best validation rewards.

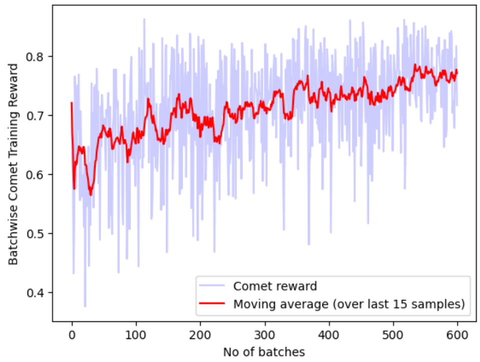
4.3.1 Training evaluation:

During training, we test model's progress in each iteration by calculating $R_{translation}$ and R_{PGL} and averaging it over the mini-batch. These plots are presented in Fig. 3 for with-context model training and Fig. 2 for the without-context case. Moreover, in every 100 iterations, we perform a validation analysis during training. We calculate the average training reward and the cross-entropy loss in the validation set. We use these scores to keep track of model's training and to choose the model checkpoint with the highest training reward on validation set.

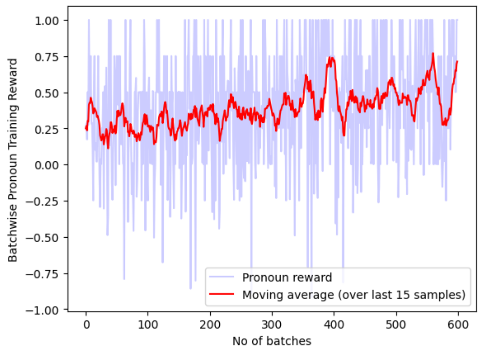
4.3.2 Testing evaluation:

We evaluate the best model checkpoint found during training based on several open-source benchmarks as listed below:

- **BLEU** (Bilingual Evaluation Understudy)
- **COMET** (Crosslingual Optimized Metric for Evaluation of Translation): We employ two versions of COMET. For direct assessment, we use the *Unbabel/wmt22-comet-da* model, which is fine-tuned on human evaluation data from the WMT22 Metrics Shared Task and name this *COMET* in the results. For quality estimation without reference translations, we use the 'wmt21-comet-qe-da' model and

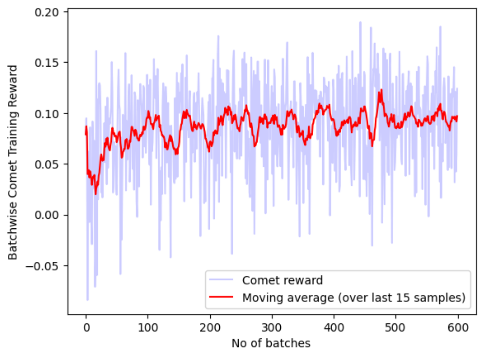


(a)

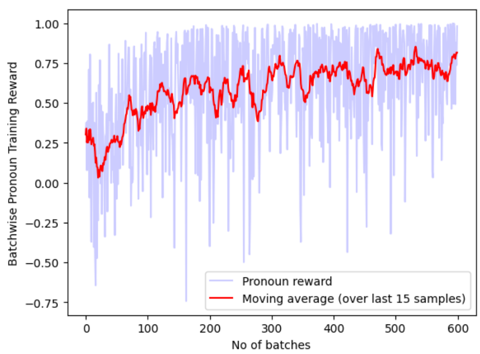


(b)

Figure 2: Batchwise training translation reward (a) and pronoun reward (b) for without-context model.



(a)



(b)

Figure 3: Batchwise training translation reward (a) and pronoun reward (b) for with-context model.

Method	En⇒De			
	COMET	BLEU	QE	PGL
NLLB 600M - DISTILLED				
NO CONTEXT				
α β				
1 $1/\#tokens$	72.88	15.2200	**0.1189**	**0.3769**
1 $1/\#avg-len$	**73.95**	**15.2630**	0.1186	0.3672
1.2 $1/\#tokens$	63.30	10.8723	0.0069	0.3425
1.2 $1/\#avg-len$	66.03	11.5783	0.0598	0.3370
WITH CONTEXT				
α β				
1 $1/\#tokens$	78.66	21.2250	0.0127	0.8270
1 $1/\#avg-len$	78.98	21.9542	0.0121	**0.8543**
1.2 $1/\#tokens$	**81.92**	**26.9574**	0.0362	0.4183
1.2 $1/\#avg-len$	66.03	11.5783	**0.0598**	0.337

Table 2: Translation evaluation on test set for En⇒De direction under various combinations of α and β, using WMT21-COMET-QE-DA as reward model. $\#tokens$ refers to number of tokens in the respective sentence, $\#avg-len$ refers to the calculated average token length across the source side dataset, calculated to be approximately 30. QE ($R_{translation}$) and PGL (R_{PGL}) refer to the respective average reward calculated on the test set.

Method	En⇒De				
	Loss	COMET	BLEU	QE	PGL
NLLB 600M - DISTILLED					
NO CONTEXT					
SFT	**1.0507**	56.68	7.8904	-0.1154	0.0902
ONLY PGL REWARD	2.2632	50.30	1.8290	-0.1534	**0.8998**
ONLY QE REWARD	2.2521	58.30	2.6723	0.048	0.3798
PRONMT ($\alpha*$,$\beta*$)	2.1593	**73.95**	**15.2630**	**0.1186**	0.3672
WITH CONTEXT					
BASELINE	6.807	66.03	11.5783	0.0598	0.337
SFT	1.2814	81.19	25.568	0.0258	0.1406
ONLY PGL REWARD	2.2340	16.79	0.0703	-0.6024	**0.9675**
ONLY QE REWARD	3.606	80.06	21.5588	0.0303	0.2262
PRONMT ($\alpha*$,$\beta*$)	**1.1486**	**81.92**	**26.9574**	**0.0362**	0.4183

Table 3: Translation performance comparison of best combination with reward baselines, using WMT21-COMET-QE-DA as reward model for EN⇒DE direction. QE ($R_{translation}$) and PGL (R_{PGL}) refer to the respective average reward calculated on the test set. $\alpha*$ and $\beta*$ refer to the hyperparameter configurations with highest COMET and BLEU scores (highlighted in bold in Table 2).

present it under the *QE* column in the results. This QE model predicts the quality of a translation based solely on the source sentence and the translation hypothesis.

By combining these metrics, we obtain a comprehensive evaluation of the model's performance in terms of accuracy, fluency, and alignment with human judgments.

5 Results

We present the translation assessment of NLLB 600M - distilled when trained on different chosen configurations in Table 3. The evaluation results reveal that incorporating document-level context significantly enhances both pronoun-specific and general translation quality, as demonstrated by the superior performance of context-aware models across all metrics. For instance, in the EN→DE direction, the context-aware ProNMT achieved a COMET score of 81.92 and a BLEU score of 26.95, compared to 73.95 and 15.2630, respectively, for the context-agnostic model. Pronoun-specific rewards, particularly those leveraging Pronoun Generation Likelihood (PGL), led to notable improvements in pronoun handling, with the context-aware model achieving a PGL score of 0.4183 versus 0.3672 for the baseline. However, models trained solely with PGL rewards underperformed on overall translation metrics, highlighting the importance of balancing PGL with Quality Estimation (QE) rewards. The combined use of QE and PGL rewards, optimized with appropriate weight configurations, yielded the best results, as evidenced by consistent improvements in batch-wise rewards during training. Context-agnostic models struggled to resolve inter-sentential dependencies, further underscoring the necessity of leveraging document-level context for coherent and accurate translations. We observed two key trends in the evaluation results: first, the inclusion of document-level context significantly enhances both pronoun-specific and overall translation quality, as evidenced by the context-aware ProNMT achieving higher scores across metrics such as COMET (81.92 vs. 73.95) and BLEU (26.95 vs. 15.2630) compared to its context-agnostic counterpart.

Second, ProNMT's ability to jointly optimize QE and pronoun-specific rewards led to consistent improvements in external metrics like COMET and BLEU, particularly when both reward components ($\alpha \neq 0, \beta \neq 0$) were employed. This balance of rewards proved crucial for achieving concurrent gains in general translation quality and accurate handling of linguistically challenging pronoun translations.

6 Conclusion

In this paper, we introduced ProNMT, a novel framework designed to address the longstanding challenges of pronoun translation in Neural Machine Translation (NMT) systems. By leveraging Quality Estimation (QE) models and the Pronoun Generation Likelihood-Based Feedback mechanism, ProNMT effectively improves both pronoun-specific and overall translation quality without the need for extensive human annotations. Our method uniquely integrates QE-based evaluations with pronoun-specific rewards, guiding iterative fine-tuning processes that are scalable, efficient, and context-aware.

Extensive experimental evaluations demonstrated that ProNMT consistently outperforms baseline systems across multiple metrics, including COMET, BLEU and QE models. Importantly, incorporating document-level context significantly enhanced the handling of linguistically complex elements, such as pronouns, while maintaining high performance on general translation tasks. These results validate the framework's ability to address both inter-sentential dependencies and broader document coherence in machine translation.

References

Ramakrishna Appicharla, Baban Gain, Santanu Pal, and Asif Ekbal. 2023. A case study on context encoding in multi-encoder based document-level neural machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 160–172, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Ramakrishna Appicharla, Baban Gain, Santanu Pal, Asif Ekbal, and Pushpak Bhattacharyya. 2024. A case study on context-aware neural machine translation with multi-task learning. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 246–257, Sheffield, UK. European Association for Machine Translation (EAMT).

Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8164–8180, Mexico City, Mexico. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 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. Association for Computational Linguistics.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018a. Can neural machine translation be improved with user feedback? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 92–105, New Orleans - Louisiana. Association for Computational Linguistics.

Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018b. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Melbourne, Australia. Association for Computational Linguistics.

Bei Li, 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. 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 Neu-

Dataset	Mean	Median	Variance	Standard Deviation	Max Length	Min Length
Contrapro	12.533	11.0	53.031	7.282	67	2
Europarl	33.378	29.0	404.219	20.105	212	1

Table 4: Statistical properties of the Contrapro and Europarl datasets.

ral Machine Translation. In *WMT 2018*, Brussels, Belgium. Association for Computational Linguistics.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. *Preprint*, arXiv:2210.01241.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Karin Sim Smith. 2017. On integrating discourse in machine translation. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 110–121, Copenhagen, Denmark. Association for Computational Linguistics.

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.

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.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

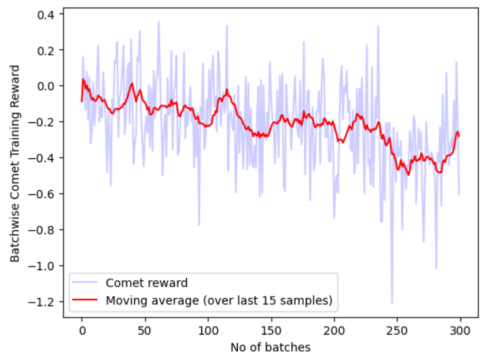
Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational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

Longyue Wang, Zhaopeng Tu, Xing Wang, Li Ding, Liang Ding, and Shuming Shi. 2020. Tencent AI lab machine translation systems for WMT20 chat translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 483–491, Online. Association for Computational Linguistics.

A Appendix

A.1 Training on all German pronouns

In our initial experiments, we tried to incorporate all German pronouns into our training framework. We considered the following consolidated list of pronouns: ['mein', 'uns', 'euer', 'ihnen', 'der', 'ihm', 'die', 'euch', 'diesen', 'unser', 'dem', 'denen', 'dieses', 'meinem', 'den', 'diese', 'du', 'seiner', 'meines', 'das', 'ich', 'deiner', 'dich', 'dir', 'meiner', 'meinen', 'es', 'meine', 'wir', 'sein', 'ihn', 'deren', 'diesem', 'sie', 'dessen', 'dieser', 'mich', 'ihr', 'mir', 'derer/deren', 'dein', 'ihrer', 'er']. Upon running ProNMT on this pronoun list, we get the results presented in Fig. 5. We attribute the noisy nature of the training curves to the noise introduced by the PGL reward and the possible ambiguity associated with translation in this scenario (Müller et al., 2018).



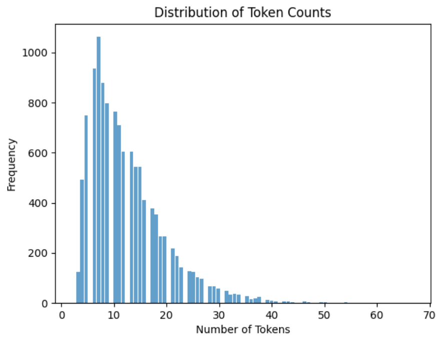
(a)

(b)

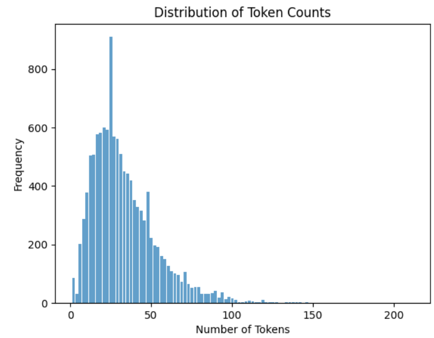
Figure 4: Batchwise training translation reward (a) and pronoun reward (b) for with-context model when no german pronouns are filtered.

A.2 Source side token distribution in datasets

In this section, we will contrast the Europarl and Contrapro datasets with respect to the source side sentence token count distribution. Performing a basic statistical analysis, we get the results as presented in Table 4 and Figure 5.



(a)



(b)

Figure 5: Token counts distribution for (a) Contrapro and (b) filtered Europarl dataset.

We hypothesize that the lower mean length and left-shifted distribution for the contrapro data set made the Quality estimation model scores saturated from the beginning of training, giving the model less room to learn. This was the reason we chose our own filtered Europarl dataset for training, testing, and validation.

B Limitations

We identify the following limitations in our work:

- The experiments were conducted only for NLLB 600M distilled variant. To assess the robustness of our framework across MT models, we can expand the scope of our chosen models.
- The current framework only accommodates the translation of the pronoun "it". We can extend the framework to include non-overlapping pronouns, i.e., pronouns whose translation is not ambiguous.
- We could not perform hyperparameter tuning for the SFT model training. The hyperparameters presented in Table 1 are default hyperparameters.
- We only consider the EN→DE direction in our experiments as it is considered to be a more difficult task than the opposite direction in MT.