Progressive Translation: Improving Domain Robustness of Neural Machine Translation with Intermediate Sequences*

Chaojun Wang[1]   **Yang Liu**[2]   **Wai Lam**[1]

[1]The Chinese University of Hong Kong
[2]Microsoft Cognitive Services Research
cj.wang@link.cuhk.edu.hk

Abstract

Previous studies show that intermediate supervision signals benefit various Natural Language Processing tasks. However, it is not clear whether there exist intermediate signals that benefit Neural Machine Translation (NMT). Borrowing techniques from Statistical Machine Translation, we propose intermediate signals which are intermediate sequences from the "source-like" structure to the "target-like" structure. Such intermediate sequences introduce an inductive bias that reflects a domain-agnostic principle of translation, which reduces spurious correlations that are harmful to out-of-domain generalisation. Furthermore, we introduce a full-permutation multi-task learning to alleviate the spurious causal relations from intermediate sequences to the target, which results from *exposure bias*. The Minimum Bayes Risk decoding algorithm is used to pick the best candidate translation from all permutations to further improve the performance. Experiments show that the introduced intermediate signals can effectively improve the domain robustness of NMT and reduces the amount of hallucinations on out-of-domain translation. Further analysis shows that our methods are especially promising in low-resource scenarios.

1 Introduction

A spectrum of studies recently arose in Natural Language Processing (NLP), which incorporates intermediate supervision signals into the model by simply converting the intermediate signals into textual sequences and prepending or appending these sequences to the output sequence. It benefits tasks such as math word problems (Wei et al., 2022), commonsense reasoning (Liu et al., 2022), programs execution (Nye et al., 2022), summarisation (Narayan et al., 2021), etc. This trend further
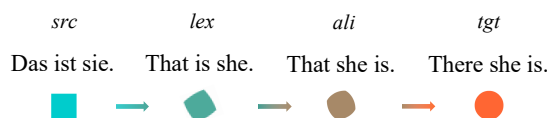

Figure 1: An illustration of the transformation from a source sentence to the target translation and its analogy with vision. *src*: source; *tgt*: target; *lex*: word-by-word translation; *ali*: reorders *lex* monotonically based on word alignments.

triggered the collection of a new dataset with intermediate results (Lewkowycz et al., 2022) and corresponding theoretical analysis (Wies et al., 2022). Intermediate supervision signals show consistent benefits to these various sequence generation tasks and Neural Machine Translation (NMT) is a basic and typical sequence generation task in the NLP community. However, it remains an open question whether and how intermediate signals can be defined and leveraged for NMT.

Meanwhile, previous studies (Koehn and Knowles, 2017; Müller et al., 2020) found that NMT suffers from poor domain robustness, i.e. the generalisation ability to unseen domains. Such an ability not only has theoretical meaning, but also has practical value since: 1) the target domain(s) may be unknown when a system is built; 2) some language pairs may only have training data for limited domains. Since the recent study (Wei et al., 2022) in intermediate supervision signals showed a benefit of such signals on out-of-domain generalisation, we expect intermediate signals may benefit domain robustness in NMT.

Different from math problem-solving tasks, machine translation tasks do not have explicit intermediate results to serve as the intermediate signals. A recent work (Voita et al., 2021b) found that NMT acquires the three core SMT competencies, target-side language modelling, lexical translation and reordering in order during the course of the training. Inspired by this work, we borrow tech-

niques in SMT to produce intermediate sequences as the intermediate signals for NMT. Specifically, we first obtain the word alignments for the parallel corpus and use it to produce the word-for-word translations (*lex*) and the aligned word-for-word translations (*ali*) to resemble the lexical translation and reordering competencies in SMT. As shown in Figure 1, the intermediate sequences resemble structurally approaching the target from the source progressively, which shares a similar spirit of how humans do translation or reasoning about translation step by step, thus named Progressive Translation.

Our intuition is that these intermediate sequences inject an inductive bias about a domain-agnostic principle of the transformation between two languages, i.e. word-for-word mapping, then reordering, and finally refinement. Such a bias limits the learning flexibility of the model but prevents the model from building up some spurious correlations (Arjovsky et al., 2019) which harm out-of-domain performance.

However, previous works have shown that NMT is prone to overly relying on the target history (Wang and Sennrich, 2020; Voita et al., 2021a), which is partially correlated with *exposure bias* (Ranzato et al., 2016) (a mismatch between training and inference), especially under domain-shift. Simply prepending these introduced intermediate sequences to the target would introduce spurious causal relationships from the intermediate sequences to the target. As a result, these intermediate sequences would potentially mislead the model about the prediction of the target, due to erroneous intermediate sequences during inference. To alleviate this spurious causal relationship, we introduce the full-permutation multi-task learning framework, where the target and intermediate sequences are fully permuted. The Minimum Bayes Risk (Goel and Byrne, 2000) decoding algorithm is used to select a *consensus* translation from all permutations to further improve the performance.

We first test our proposed framework on IWSLT'14 German→English and find that the proposed intermediate sequence can improve the domain robustness of NMT. The permutation multi-task learning is important for the intermediate sequence which is prone to erroneous during inference. To examine the generality of our methods, we conduct experiments on another two domain-robustness datasets in NMT,

OPUS German→English and a low resource German→Romansh scenario. Our methods show consistent out-of-domain improvement over these two datasets.
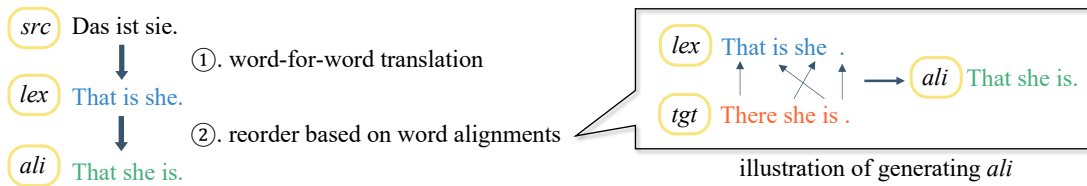
Moreover, previous works (Müller et al., 2020; Wang and Sennrich, 2020) found that hallucinated translations are more pronounced in out-of-domain setting. Such translations are fluent but completely unrelated to the input, and they may cause more serious problems in practical use due to their misleading nature. Therefore, we manually evaluate the proportion of hallucinations. Results show that our methods substantially reduce the amount of hallucinations in out-of-domain translation. Finally, since the corpus size in the main experiments is relatively small, we investigate the effectiveness of our methods when scaling up the corpus sizes. Results show that our methods are especially effective under the low-resource scenarios.

2 Related Work

Intermediate Supervision Signals. Some existing works in the broader NLP community try to incorporate intermediate sequences into the model. We take two typical examples of them to better distinguish our work from other works. Narayan et al. (2021) uses an entity chain as the intermediate sequence for summarisation. Wei et al. (2022) produces intermediate sequences resembling the deliberation process of humans. Similar to Narayan et al. (2021), Progressive Translation (PT) augments data for the whole training set and the intermediate sequences are not limited to literally understandable sequences. Similar to Wei et al. (2022), sequences augmented by PT resemble approaching the output from the input.

Data Augmentation of Domain Robustness in NMT. Existing works in data augmentation try to improve the domain robustness of NMT by introducing more diverse synthetic training examples (Ng et al., 2020) or auxiliary tasks where the target history is less informative (Sánchez-Cartagena et al., 2021) named MTL-DA framework. The main difference between our PT framework and the MTL-DA framework is that the MTL-DA framework treats each target-side sequence as an independent task conditioned on the source, whereas PT also encourages the model to learn the transformational relations between any pair of target-side sequences, which may help the model to generalise better across domains.

Data Augmentation:

① . word-for-word translation

② . reorder based on word alignments

illustration of generating *ali*

Multi-task Learning:

<123> is a control token indicating the order of three sequences. 1: *lex*; 2: ali; 3: *tgt*, then <123> is for the task where the target is in order of *lex, ali and tgt*. <lex>, <ali>, <tgt> is the special tokens prepended to lex, ali, tgt separately.

Old training pair:

Source: Das ist sie.
Target: There she is.

New training pairs:

Source: <123> Das ist sie. Target: <lex> That is she. <ali> That she is. <tgt> There she is.

Source: <321> Das ist sie. Target: <tgt> There she is. <ali> That she is. <lex> That is she.
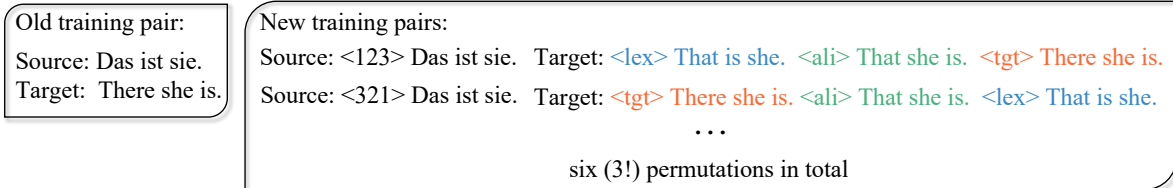
· · ·

six (3!) permutations in total

Figure 2: An illustration of the proposed intermediate sequences and multi-task learning framework. src: source.

Statistical Machine Translation in NMT. The intermediate sequences of PT are produced using the word alignments and reordering components in Statistical Machine Translation (SMT). There are works on improving NMT with SMT features and techniques (He et al., 2016; Chen et al., 2016; Du and Way, 2017; Zhao et al., 2018). However, these works either modify the architecture of the neural network or require more than one model to produce the translation (e.g. a rule-based pre-ordering model and a NMT model etc.). To the best of our knowledge, we are the first to incorporate features from SMT into NMT by converting the features into textual sequences and prepending these to the target without requiring extra models or modifying the neural architecture.

3 Approach

3.1 Intermediate Sequences

The traditional SMT decomposes the translation task into distinct components where some features could potentially be the intermediate supervision signals. More recently, Voita et al. (2021b) found that NMT acquires the three core SMT competencies, i.e. target-side language modelling, lexical translation and reordering, in order during the course of training. Inspired by this work, we produce word-for-word translations and aligned word-for-word translations as the intermediate sequences to resemble the lexical translation and reordering components separately using the word alignments component in SMT.

As shown in Figure 2 Data Augmentation part, for each source-target parallel sequence in the training corpus, we augment their target sequences with two extra intermediate sequences, *lex* and *ali*. The two intermediate sequences are prepended to the target to form an augmented target.

lex: The source sequence is word-for-word translated based on a bilingual lexicon obtained from the parallel training corpus. Tokens that are not in the lexicon are copied into *lex*.

ali: *lex* is reordered so that the word alignments from the target to *lex* is monotonic. The word alignments used here are target-to-source alignments because it is equivalent to the target-to-*lex* alignments since *lex* is word-for-word mapped from the source. The words in the target which is assigned to "NULL" are omitted during reordering.

lex, *ali* and target (*tgt*) are prefixed with a special token separately for extracting the corresponding sequence from the predicted output. The one-to-many (both source-to-target and target-to-source) word alignments are obtained with *mgiza++* (Gao and Vogel, 2008; Och and Ney, 2003)[1], a SMT word alignments tool, on the **in-domain** training corpus, following the default parameter provided in *train-model.perl* by Moses (Koehn et al., 2007)[2]. The one-to-one word alignments are built by computing the intersection between the one-to-many word alignments in both directions. The bilingual lexicon is obtained by associating each source word

---

[1] https://github.com/moses-smt/mgiza
[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/train-model.perl
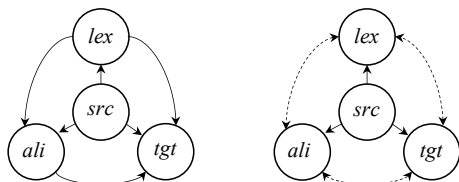
Figure 3: Causal graphs for the source and three target-side sequences. Solid arrow denotes casual dependence and dashed arrow represents the statistical correlation between two variables. Left: relations if we simply prepend *lex* and *ali* to the target. Right: relations after full-permutation multi-task learning.

to the target word it is most frequently aligned within the one-to-one word alignments.

The learning of word alignments and transformations of *lex* and *ali* are at the word level. The BPE (Sennrich et al., 2016) word segmentation is trained on *src-tgt* parallel data as normal and applied to both source-target parallel sequences and intermediate sequences (the target-language vocabulary is applied to split the words in the intermediate sequences).

We expect that the introduced intermediate sequences would benefit the domain robustness of NMT. Because the proposed intermediate sequences serve as a supervision signal to provide the model with an explicit path for learning the transformational relations from source to target. Such signals inject an inductive bias about one kind of domain-agnostic principle of the transformation between two languages, i.e. word-for-word mapping, then reordering, finally refinement. This injected bias limits the learning flexibility of the neural model but prevents the model from building up some spurious correlations which harm out-of-domain performance.

3.2 Spurious Causality Relationship

To introduce these intermediate sequences as intermediate supervision signals to the model, we prepend them to the output sequence in training. However, simply prepending these produced intermediate sequences to the target would potentially introduce spurious causality relationships from pre-sequence to post-sequence. For example, prepending *lex*, *ali* to the target would introduce the causal relationships of $lex \rightarrow ali \rightarrow tgt$. These are spurious causality relationships because the model is highly unlikely to get the gold-standard pre-sequences (*lex* or *ali*) as in the training during inference, especially under the domain-shift where the performance is

relatively poor. Therefore, the model should learn that source (input) is the only reliable information for any target-side sequences. Note that such spurious causality relationship in principle results from a mismatch between training and inference of the standard training-inference paradigm of NMT, which is termed *exposure bias* by the community.

Intuitively, if the model could predict the target-side sequences in any order, then the causality relationship between target-side sequences should be reduced. Therefore, we propose to fully permute the target-side sequences, i.e. intermediate sequences (*lex* or *ali*) and the target sequence (*tgt*). Figrue 2 illustrates the training data after permutation when we prepend both *lex* and *ali* to the target. The source is prefixed with a control token for each permutation, i.e. 1: *lex*; 2: *ali*; 3: *tgt*, then <123> is the control token for the permutation where the target is in the order of *lex*, *ali* and *tgt*.

As shown in Figure 3, with the permutation, we create counterfactual data which disentangles the causal relations of $lex \rightarrow ali \rightarrow tgt$ and enhances the causal relations from source to each of these three sequences. Therefore, the full-permutation multi-task training better balances the model's reliance on the source and target history, at least on pre-sequence(s).

3.3 Minimum Bayes Risk Decoding

From our preliminary experiments, we found that various test sets prefer different generation orders of the permutation. For example, order *lex-ali-tgt* performs best on some test sets whereas *tgt-ali-lex* performs best on some other test sets. Therefore, we suspect that the translation quality would be further improved if we could dynamically select the best candidate translations from all permutations. Inspired by (Eikema and Aziz, 2021), we use Minimum Bayes Risk (MBR) decoding to select a *consensus* translation from all permutations.

MBR aims to find a translation that maximises expected utility (or minimises expected risk) over the posterior distribution. In practice, the posterior distribution is approximated by drawing a pool of samples $\mathcal{S} = (s_1, ..., s_n)$ of size n from the model:

$$y^\star = \underset{s_i \in \mathcal{S}}{\mathrm{argmax}} \frac{1}{n} \sum_{s_j=1}^{n} u\left(s_i, s_j\right) \qquad (1)$$

where u is the utility function to compute the similarity between two sequences. In our experiment,

the samples $\mathcal{S}$ are translations from all permutations.

Following Eikema and Aziz (2021), we use BEER (Stanojević and Sima'an, 2014) as the utility function, and the released toolkit[3] for MBR decoding.

4 Experiments

4.1 Dataset

We work on three datasets involving two language pairs, which were used in previous works on the domain robustness in NMT (Sánchez-Cartagena et al., 2021; Ng et al., 2020).

IWSLT'14 DE→EN IWSLT'14 (Cettolo et al., 2014) German→English (DE→EN) is a commonly used small-scale dataset in NMT, which consists of 180 000 sentence pairs in the TED talk domain. Following Sánchez-Cartagena et al. (2021), the validation and in-domain (ID) testing sets are *tst2013* and *tst2014* separately; and out-of-domain (OOD) test sets consist of *IT*, *law* and *medical* domains from OPUS (Lison and Tiedemann, 2016) collected by Müller et al. (2020)[4].

OPUS DE→EN & Allegra DE→RM are two benchmarks of domain-robustness NMT released by Müller et al. (2020). OPUS comprises five domains: *medical*, *IT*, *law*, *koran* and *subtitles*. Following Ng et al. (2020), we use *medical* as ID for training (which consists of 600 000 parallel sentences) and validation and the rest of four domains as OOD test sets. Allegra (Scherrer and Cartoni, 2012) German→Romansh (DE→RM) has 100 000 sentence pairs in *law* domain. The test OOD domain is *blogs*, using data from Convivenza.

We tokenise and truecase all datasets with Moses and use shared BPE with 10 000 (on IWSLT'14) and 32 000 (on OPUS and Allegra) for word segmentation (Sennrich et al., 2016).

4.2 Models and Evaluation

All experiments are done with the Nematus toolkit (Sennrich et al., 2017) based on the Transformer architecture (Vaswani et al., 2017)[5]. The baseline is trained on the training corpus without using intermediate sequences. We follow Wang and Sennrich (2020) to set hyperparameters (see

Appendix) on three datasets. For our framework, we scale up the token batch size proportional to the length of the target for a fair comparison, e.g. if the target-side sequence is three times longer than the original target, we scale up the batch size to three times as well.[6] The performance of the original order *(lex)-(ali)-tgt* is used for validation and testing. We conduct early-stopping if the validation performance underperforms the best one over 10 times of validation in both the translation quality (BLEU) and the cross entropy loss.

We also compare to two recently proposed methods of domain robustness in NMT. SSMBA (Ng et al., 2020) generates synthetic training data by moving randomly on a data manifold with a pair of corruption and reconstruction functions. Reverse+Mono+Replace (Sánchez-Cartagena et al., 2021) (RMP) introduces three auxiliary tasks where the target history is less informative.

We report cased, detokenised BLEU (Papineni et al., 2002) with SacreBLEU (Post, 2018)[7]. Each experiment is independently run for three times, and we report the average and standard deviation to account for optimiser instability.

4.3 Results

We test our proposal mainly on IWSLT'14 DE→EN. Table 1 summarises the results. ① is the baseline system which is trained on parallel corpus only without any data augmentation. The average OOD is computed by averaging results across all OOD test sets.

Single *lex* benefits OOD whereas *ali* does not. Firstly, we simply prepend the produced intermediate sequence(s) (any one of them and both of them in the order of *lex-ali*) to the target sequence. Results show that single *lex* (②) significantly improves the OOD performance by 2.2 BLEU, at the cost of 0.9 BLEU decrease in in-domain performance. However, the introduction of *ali* deteriorates the performance on both in-domain (ID) and OOD test sets (③ and ④). We argue that this comes from the reason that the learning of generating *ali* is more difficult than generating *lex* (*ali* needs an extra reordering step and also the produced *ali* is noisy due to the word alignment errors). As a result, *ali* is more erroneous than *lex* during inference. Therefore, the generation quality of the

---

[6]Scaling up the token batch size only brings negligible improvement on the baseline.

[7]Signature: BLEU|#:1|c:mixed|e:no|tok:13a|s:exp|v:2.1.0

ID	Augmentation	In-Domain	IT	Law	Medical	average OOD
1	Transformer	$32.1_{\pm0.38}$	$14.7_{\pm0.21}$	$10.1_{\pm0.38}$	$17.0_{\pm0.25}$	$13.9_{\pm0.19}$
2	*lex+tgt*	$31.2_{\pm0.50}$	$16.6_{\pm0.26}$	$11.1_{\pm0.23}$	$20.7_{\pm0.66}$	$16.1_{\pm0.30}$
3	*ali+tgt*	$25.8_{\pm3.57}$	$14.4_{\pm2.54}$	$4.5_{\pm6.00}$	$17.9_{\pm1.32}$	$12.2_{\pm3.25}$
4	*lex+ali+tgt*	$25.5_{\pm7.82}$	$9.4_{\pm1.14}$	$3.1_{\pm2.31}$	$11.3_{\pm6.70}$	$7.9_{\pm1.71}$
5	2 + permu	$30.1_{\pm1.55}$	$15.5_{\pm0.50}$	$7.2_{\pm5.48}$	$19.0_{\pm1.08}$	$13.9_{\pm2.18}$
6	3 + permu	$30.6_{\pm0.30}$	$16.9_{\pm1.00}$	$10.8_{\pm0.40}$	$19.9_{\pm0.60}$	$15.9_{\pm0.53}$
7	4 + permu	$29.9_{\pm0.32}$	$18.2_{\pm0.89}$	$10.8_{\pm0.10}$	$20.7_{\pm0.40}$	$16.6_{\pm0.37}$
8	7 + MBR	$30.5_{\pm0.21}$	$17.7_{\pm0.72}$	$11.8_{\pm0.1}$	$21.6_{\pm0.49}$	$17.0_{\pm0.35}$

Table 1: Average BLEU (↑) and standard deviation of ablation results on in-domain and out-of-domain test sets on IWSLT'14 DE→EN. permu: permutation.

target deteriorates due to its causal dependency on *ali*.

***ali* benefits OOD with the support of permutation multi-task learning.** We try to alleviate the problem by introducing the permutation multi-task learning on top of ②∼④. Results show that the permutation successfully alleviates the deterioration of introducing *ali*, bringing positive results for both ID and OOD (③→⑥, ④→⑦). With the permutation, a single *ali* intermediate sequence (⑥) can improve OOD over the baseline by 2 BLEU and the combination of *lex* and *ali* (⑦) bring further improvement on OOD over single *lex* (②) or single *ali* (⑥) by 0.5 and 0.7 BLEU respectively. The permutation shows a negative effect on single *lex* (②→⑤). Because the *lex* is very easy to learn, few error would occur when predicting *lex*. Therefore, permutation is not effective and even has negative effects as it makes the neural model hard to focus on learning the task of *lex-tgt*, leading to inferior performance.

MBR decoding brings further improvement. For the *lex, ali, tgt* with permutation, there are six permutations in total. We dynamically select a *consensus* translation over each input data by performing MBR decoding over translation from all permutations. Results show MBR (⑦→⑧) could further improve the OOD and ID performances by 0.4 and 0.6 BLEU respectively, and outperforms baseline OOD by 3.1 BLEU at the cost of 1.6 BLEU decrease in ID.

Results on other datasets and comparison with existing methods. As ⑧ achieves the highest OOD performance and ② achieves relatively high OOD and ID performance with simpler techniques, we name ⑧ as PT_{full} and ② as PT_{simple} and evaluate these two methods on another two domain-robustness datasets (OPUS DE→EN and Allegra DE→RM). Table 2 lists the results.

Baselines (Transformer) in cited works (RMP and SSMBA) are trained under inappropriate hyperparameters, e.g. on IWSLT'14, the cited works uses default hyperparameters for the WMT dataset (more than 10 times larger than IWSLT'14). To enable better comparison by other researchers, we train the Transformer with the appropriate hyperparameters provided by Wang and Sennrich (2020) to build strong baselines, which outperform those in the cited works. We re-implement the other two DA methods based on our baseline for comparison.

Results show that both PT_{simple} and PT_{full} perform most effectively on IWSLT'14 OOD, surpassing the existing methods by 0.7-2.3 BLEU. On the other two new datasets, PT_{simple} and PT_{full} show consistent OOD improvement, outperforming our baseline (Transformer) by 1.1-1.6 BLEU and 1.1-1.2 BLEU on OPUS and DE→RM dataset respectively. The ID performance of PT_{simple} and PT_{full} on these two datasets is less affected than on IWSLT'14, at the cost of 0.3-0.4 BLUE decrease on OPUS and even no decrease on the Allegra DE→RM.

PT_{full} significantly outperforms PT_{simple} OOD on OPUS DE→EN and they show negligible ID differences. For Allegra DE→RM, PT_{simple} and PT_{full} shows similar OOD and ID performance.

5 Analysis

BLEU score indicates that the proposed methods can improve domain robustness. In this section, we investigate the reduction of hallucinations and performance on larger datasets of our methods.

5.1 Hallucinations

Hallucinations are more pronounced in out-of-domain translation, and their misleading nature makes them particularly problematic. Therefore, many works have been conducted on hallucinations,

augmentation	IWSLT'14		OPUS		DE→RM	
	in-domain	average OOD	in-domain	average OOD	in-domain	average OOD
Results reported by Sánchez-Cartagena et al. (2021):						
Transformer	$30.0_{\pm0.10}$	$8.3_{\pm0.85}$	-	-	-	-
RMP	$31.4_{\pm0.30}$	$11.8_{\pm0.48}$	-	-	-	-
Results reported by Ng et al. (2020):						
Transformer	-	-	57.0	10.2	51.5	12.2
SSMBA	-	-	54.9	10.7	52.0	14.7
Our experiments:						
Transformer	$32.1_{\pm0.38}$	$13.9_{\pm0.19}$	$58.8_{\pm0.38}$	$11.0_{\pm0.22}$	$54.4_{\pm0.25}$	$19.2_{\pm0.23}$
SSMBA	$31.9_{\pm0.15}$	$15.4_{\pm0.10}$	$58.4_{\pm0.20}$	$12.1_{\pm0.21}$	$54.7_{\pm0.20}$	$20.4_{\pm0.15}$
RMP	$32.2_{\pm0.06}$	$14.7_{\pm0.17}$	$59.2_{\pm0.25}$	$12.6_{\pm0.41}$	$55.1_{\pm0.21}$	$21.5_{\pm0.23}$
PT_{simple}	$31.2_{\pm0.50}$	$16.1_{\pm0.30}$	$58.5_{\pm0.64}$	$12.1_{\pm0.18}$	$54.6_{\pm0.12}$	$20.3_{\pm0.31}$
PT_{full}	$30.5_{\pm0.21}$	$17.0_{\pm0.35}$	$58.4_{\pm0.12}$	$12.6_{\pm0.10}$	$54.4_{\pm0.21}$	$20.4_{\pm0.51}$

Table 2: Average BLEU (↑) and standard deviation on in-domain and out-of-domain test sets for models trained on IWSLT'14 DE→EN, OPUS DE→EN and Allegra DE→RM. PT_{simple}: method ② in Table 1; PT_{full}: method ⑧ in Table 1; RMP: Reverse+Mono+Replace

involving detection of hallucinations (Zhou et al., 2021; Guerreiro et al., 2022; Dale et al., 2022), exploration of the causes of hallucinations (Raunak et al., 2021; Yan et al., 2022), and finding solutions for hallucinations (Miao et al., 2021; Müller and Sennrich, 2021) etc.

To test our methods for reducing the hallucinations under domain shift, we manually evaluate the proportion of hallucinations on IWSLT'14 and OPUS (DE→EN) OOD test sets. We follow the definition and evaluation by Müller et al. (2020), considering a translation as a hallucination if it is **(partially) fluent** and its content is not related to the source **(inadequate)**. We report the proportion of such hallucinations in each system.

The manual evaluation is performed by two students who have completed an English-medium university program. We collect ∼3000 annotations for 10 configurations. We ask annotators to evaluate translations according to fluency and adequacy. For fluency, the annotator classifies a translation as fluent, partially fluent or not fluent; for adequacy, as adequate, partially adequate or inadequate. We report the kappa coefficient (K) (Carletta, 1996) for inter-annotator and intra-annotator agreement in Table 3, and assess statistical significance with Fisher's exact test (two-tailed).

Table 4 shows the results of human evaluation. All of the DA methods significantly decrease the proportion of hallucinations by 2%-6% on IWSLT'14 and by 9%-11% on OPUS, with the increase in BLEU. Note that the two metrics do not correlate perfectly: for example, PT_{full} has a higher BLEU than PT_{simple} but PT_{simple} has a

	inter-annotator			intra-annotator		
annotation	$P(A)$	$P(E)$	K	$P(A)$	$P(E)$	K
fluency	0.52	0.31	0.30	0.84	0.39	0.73
adequacy	0.68	0.38	0.48	0.88	0.38	0.81

Table 3: Inter-annotator (N=300) and intra-annotator agreement (N=150) of manual evaluation.

similar or even lower proportion of hallucinations than PT_{full}. This indicates that PT_{full} improves translation quality in other aspects.

	% hallucinations (BLEU)	
Augmentation	IWSLT'14	OPUS
Transformer	11% (13.9)	39% (11.0)
RMP	9% (14.7)	30% (12.6)
SSMBA	6% (15.4)	28% (12.1)
PT_{simple}	5% (16.1)	28% (12.1)
PT_{full}	7% (17.0)	30% (12.6)

Table 4: Proportion of hallucinations (↓) and BLEU (↑) on out-of-domain test sets over IWSLT'14 and OPUS (DE→EN).

5.2 Tendency by scaling up the corpus size

Since the size of the training corpus in the previous experiments ranges from 0.1M to 0.6M (million) samples, which is a low-resource setting for NMT, here we investigate the performance of our methods when scaling up the corpus size. We use *subtitles* domain from OPUS as the in-domain training data (because it has around 20M sentence pairs) and the rest four domains as the OOD test sets. We use the first 0.2M, 2M and 20M samples in the corpus as the training data separately. We follow
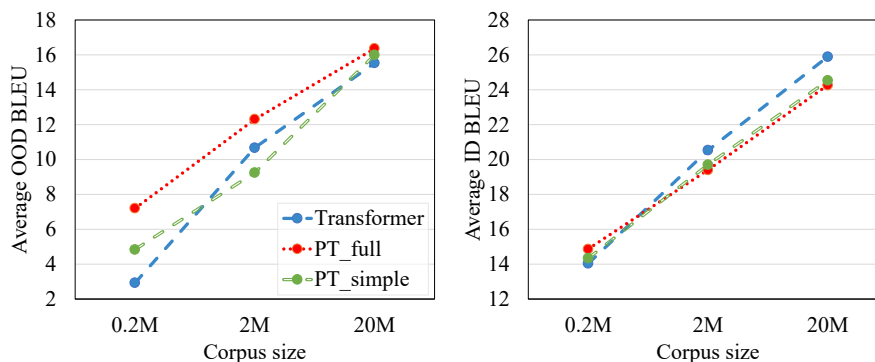
Figure 4: Average BLEU (↑) on in-domain and out-of-domain test sets for models trained on OPUS DE→EN (*subtitles*) with various sizes of the training corpus.

the same data preprocessing as for OPUS (*medical*). The hyperparameters for training the model are the same as those for IWSLT'14 when the corpus size is 0.2M and those for OPUS (*medical*) when the corpus size is 2M. For the corpus size of 20M, we increase the token batch size to 16384 instead of 4096 and keep the rest of the hyperparameters the same as for the 2M corpus size. Similarly, each experiment is independently run for three times and we report the average result.

Results are shown in Figure 4. As expected, increasing the corpus size (0.2M-20M) improves both ID and OOD performance for all systems. When the corpus size is small (0.2M), PT_{full} (red line) shows a considerable improvement in OOD over the baseline (blue line) by 4.3 BLEU and even slightly benefits ID, surpassing the baseline by around 0.9 BLEU. However, scaling up the corpus size (0.2M-20M) narrows the gap of OOD improvement (4.3-0.9 BLEU) between the baseline and PT_{full}, and widens the ID deterioration from +0.9 to -1.6 BLEU.

In general, PT_{simple} (green line) follows a similar tendency as PT_{full}, compared to the baseline. However, PT_{simple} underperforms the baseline at the corpus size of 2M. By a close inspection, we found that the training of PT_{simple} is relatively unstable. The standard deviations of PT_{simple} for OOD are 1.38, 2.49 and 0.24 on 0.2M, 2M and 20M corpus size respectively, whereas the standard deviations of PT_{full} are 0.47, 0.27 and 0.52 respectively. This indicates that the training of PT_{simple} is less stable than PT_{full} when the corpus size is 0.2M-2M. The better stability of PT_{full} may come from its permutation multi-task learning mechanism.

PT_{simple} always underperforms PT_{full} on OOD for any corpus size. PT_{simple} shows slightly better ID performance than PT_{full} when the corpus size is large (2M-20M) but underperforms PT_{full} on ID performance in low resource setting where the corpus size is 0.2M.

6 Conclusion

Our results show that our introduced intermediate signals effectively improve the OOD performance of NMT. Intermediate sequence *lex* can benefit OOD by simply prepending it to the target. *ali* is more likely to be erroneous during inference than *lex*, which results in degenerated target due to the spurious causal relationship. Our proposed permutation multi-task learning successfully alleviates the problem and manifests the effectiveness of *ali*. Experiments also confirm that the MBR algorithm can further improve the performance by dynamically selecting a *consensus* translation from all permutations. The human evaluation shows that the proposed methods substantially reduce the number of hallucinations of the out-of-domain translation. Experiments on the larger corpus sizes indicate that our methods are especially promising in the low-resource scenarios.

Our work is the first attempt to complete the puzzle of the study of intermediate signals in NMT, and two new ideas may benefit this study in other areas: 1) thinking intermediate signals from the intermediate structures between the transformation from the input to the output; 2) the permutation multi-task learning, instead of only pre/appending intermediate sequences to the output sequence. The permutation multi-task learning + MBR decoding framework is also a potential solution for any multi-pass generation tasks (e.g. speech translation), which suffer from the error propagation problem. The problem is alleviated with the permutation which

disentangles causal relations between intermediate and final results. Finally, our work provides a new perspective of data augmentation in NMT, i.e. augmenting data by introducing extra sequences instead of directly modifying the source or target.

Acknowledgements

We thank Tianzi Wang for insightful suggestions and pointers to Minimum Bayes Risk decoding algorithm and Zengrui Jin for computing resources.

7 Limitations

The way we use the intermediate sequences is to concatenate new sequences and the target sequence as the new target. As a result, the length of the target increases linearly with the number of intermediate sequences introduced, which increases the cost of inference. In the meantime, Minimum Bayes Risk decoding needs to do prediction multiple times under different control tasks, which further increases the computational cost. However, there are potential solutions to compromise between the computational cost and quality, e.g. learning a student model by distilling the domain-robust knowledge from Progressive Translation.

8 Ethics Statement

The datasets used in the experiments are all well-known machine translation datasets and publicity available. Data preprocessing does not involve any external textual resources. Intermediate sequences generated in our data augmentation method are new symbolic combinations of the tokens in the target language. However, the final output of the model is the *tgt* sequence which is the same as the target sequence in the original training set. Therefore, we would not expect the model trained with our data augmentation method would produce more harmful biases. Finally, we declare that any biases or offensive contexts generated from the model do not reflect the views or values of the authors.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 2–17, Lake Tahoe, California.

Wenhu Chen, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter. 2016. Guided alignment training for topic-aware neural machine translation. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, pages 121–134, Austin, TX, USA. The Association for Machine Translation in the Americas.

David Dale, Elena Voita, Loïc Barrault, and Marta R. Costa-jussà. 2022. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better.

Jinhua Du and Andy Way. 2017. Pre-reordering for neural machine translation: Helpful or harmful? *Prague Bulletin of Mathematical Linguistics*, (108):171–181.

Bryan Eikema and Wilker Aziz. 2021. Sampling-based minimum bayes risk decoding for neural machine translation.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio. Association for Computational Linguistics.

Vaibhava Goel and William J Byrne. 2000. Minimum bayes-risk automatic speech recognition. *Computer Speech&Language*, 14(2):115–135.

Nuno M. Guerreiro, Elena Voita, and André F. T. Martins. 2022. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation.

Wei He, Zhongjun He, Hua Wu, and Haifeng Wang. 2016. Improved neural machine translation with smt features. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

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.

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

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in neural 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 3456–3468, Online. Association for Computational Linguistics.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural 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 259–272, Online. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283, Online. Association for Computational Linguistics.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber,

David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2022. Show your work: Scratchpads for intermediate computation with language models. In *Deep Learning for Code Workshop*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2021. Rethinking data augmentation for low-resource neural machine translation: A multitask learning approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8502–8516, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yves Scherrer and Bruno Cartoni. 2012. The trilingual ALLEGRA corpus: Presentation and possible use for lexicon induction. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2890–2896, Istanbul, Turkey. European Language Resources Association (ELRA).

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages

65–68, Valencia, Spain. Association for Computational Linguistics.

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

Miloš Stanojević and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 202–206, Doha, Qatar. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021a. Analyzing the source and target contributions to predictions in neural 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 1126–1140, Online. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021b. Language modeling, lexical translation, reordering: The training process of NMT through the lens of classical SMT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8478–8491, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models.

Noam Wies, Yoav Levine, and Amnon Shashua. 2022. Sub-task decomposition enables learning in sequence to sequence tasks.

Jianhao Yan, Fandong Meng, and Jie Zhou. 2022. Probing causes of hallucinations in neural machine translations.

Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2018. Exploiting pre-ordering for neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

A Appendix

A.1 Discussion of Intermediate Sequences

lex and *ali* intermediate sequences may come from certain intermediate topological spaces between the transformation from the topological spaces of the source into the target languages. We empirically confirm that such intermediate sequences might look strange but are easier for the neural model to learn and predict, since they are structurally closer to the source. We use the standard Transformer model to learn to predict *lex*, *ali* and *tgt* (this is just the baseline) directly on IWSLT'14 dataset and report the results on both in-domain and out-of-domain test sets. Note that the gold-standard sequences of *lex* and *ali* on the out-of-domain test sets are produced on the corresponding out-of-domain training sets.

Table 5 shows that *lex* is easier to be predicted than *ali*, and *ali* is easier to be predicted than *tgt* by the NMT model, over both in-domain and out-of-domain test sets.

Domain	lex	ali	tgt
ID	$94.0_{\pm 0.20}$	$61.1_{\pm 0.12}$	$32.1_{\pm 0.38}$
OOD	$72.6_{\pm 0.60}$	$47.9_{\pm 0.48}$	$13.9_{\pm 0.19}$

Table 5: Average BLEU ($\uparrow$) and standard deviation on in-domain and out-of-domain test sets on IWSLT'14 DE$\rightarrow$EN when the target is *lex*, *ali* or *tgt* separately.

A.2 Hyperparameters

	IWSLT	OPUS/Allegra
embedding layer size		512
hidden state size		512
tie encoder decoder embeddings		yes
tie decoder embeddings		yes
loss function		per-token-cross-entropy
label smoothing		0.1
optimizer		adam
learning schedule		transformer
warmup steps	4000	6000
gradient clipping threshold	1	0
maximum sequence length		100
token batch size		4096
length normalization alpha	0.6	1
encoder depth		6
decoder depth		6
feed forward num hidden	1024	2048
number of attention heads	4	8
embedding dropout	0.3	0.1
residual dropout	0.3	0.1
relu dropout	0.3	0.1
attention weights dropout	0.3	0.1
beam size		4
validation frequency		4000 iterations

Table 6: Configurations of NMT systems over three datasets.

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

4.1

☑ B1. Did you cite the creators of artifacts you used?
4.1

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
It has been provided in the cited work.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4.1

C ☑ Did you run computational experiments?

4

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Our model is the standard Transformer model without any modification to the model's architecture.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4.3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4.1

D ☑ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

5.1

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Instructions are rather simple which could be indicated in section 5.1.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.