Revisiting Commonsense Reasoning in Machine Translation: Training, Evaluation and Challenge

Xuebo Liu[1]* **Yutong Wang**[1] **Derek F. Wong**[2] **Runzhe Zhan**[2]
Liangxuan Yu[2] **Min Zhang**[1]
[1]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
{liuxuebo,zhangmin2021}@hit.edu.cn, wangyutong200012@gmail.com
[2]NLP[2]CT Lab, Department of Computer and Information Science, University of Macau
derekfw@um.edu.mo, nlp2ct.{runzhe,liangxuan}@gmail.com

Abstract

The ability of commonsense reasoning (CR) decides whether a neural machine translation (NMT) model can move beyond pattern recognition. Despite the rapid advancement of NMT and the use of pretraining to enhance NMT models, research on CR in NMT is still in its infancy, leaving much to be explored in terms of effectively training NMT models with high CR abilities and devising accurate automatic evaluation metrics. This paper presents a comprehensive study aimed at expanding the understanding of CR in NMT. For the training, we confirm the effectiveness of incorporating pretrained knowledge into NMT models and subsequently utilizing these models as robust testbeds for investigating CR in NMT. For the evaluation, we propose a novel entity-aware evaluation method that takes into account both the NMT candidate and important entities in the candidate, which is more aligned with human judgement. Based on the strong testbed and evaluation methods, we identify challenges in training NMT models with high CR abilities and suggest directions for further unlabeled data utilization and model design. We hope that our methods and findings will contribute to advancing the research of CR in NMT. Source data, code and scripts are freely available at https://github.com/YutongWang1216/CR-NMT.

Contextless Syntactic Ambiguity	
SRC	捕获的/Hunting 是/is 猎人/hunter 。
REF	The hunter is hunting .
NMT	The hunter was captured .
PT-NMT	It was the hunter who caught it .
Contextual Syntactic Ambiguity	
SRC	手术/surgery 开刀的/operated 是/is 他父亲/his father，因为/because 他父亲/his father 得了重病/get serious illness。
REF	It was his father who underwent surgery, because his father was seriously ill.
NMT	The operation was performed by his father, who was seriously ill.
PT-NMT	The operation was performed on his father, who was seriously ill.
Lexical Ambiguity	
SRC	学校/school 规定/mandates 学生/students 上学/go to school 要/must 背/carry 书包/school bag。
REF	The school requires students to carry school bags.
NMT	Schools require students to recite school bags.
PT-NMT	The school requires students to carry their schoolbags at school.

Table 1: Translations of vanilla NMT and pretraining-based NMT (PT-NMT). Highlights denote the parts requiring commonsense knowledge for accurate translation. PT-NMT performs well in addressing both syntactic and lexical ambiguities.

1 Introduction

Commonsense reasoning (CR; Davis and Marcus, 2015) is the ability to understand and navigate the world using basic knowledge and understanding that is shared by most people. In the context of neural machine translation (NMT; Bahdanau et al., 2015; Vaswani et al., 2017; Liu et al., 2019, 2020a), CR is important because it allows the model to move beyond simply recognizing patterns in the data and instead make more informed, nuanced translations. Recent studies have witnessed the great success of adapting self-supervised pretraining to downstream language understanding and generation tasks (Devlin et al., 2019; Song et al., 2019; Floridi and Chiriatti, 2020; Ouyang et al., 2022), and one of the major ingredients is the abundant commonsense knowledge embedded in the pretrained models (Zhou et al., 2020; Tamborrino et al., 2020). As recent studies have been studying pretraining-based neural machine translation (PT-NMT) for model improvement (Conneau et al., 2020; Liu et al., 2020b), a thorough understanding of its CR ability helps to better explain the improvement and beyond.

Despite some attempts to understand CR ability of NMT from various perspectives (e.g., word

*Corresponding author

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15536–15550

sense disambiguation (Rios Gonzales et al., 2017) and pronoun resolution (Davis, 2016)), only a few studies systematically examine the ability of NMT (He et al., 2020). Furthermore, the evaluation of CR ability of NMT is under-investigated. Current evaluation methods for CR in NMT models rely on contrastive evaluation techniques (Sennrich, 2017), which do not take into account the NMT candidates, resulting in suboptimal evaluation performance. These make it difficult to conduct research on CR in NMT. Despite the difficulties, this paper aims to provide a systematic study of CR in NMT.

Training (§3) We investigate the potential benefits of utilizing pretrained knowledge for NMT training. We evaluate CR accuracy of PT-NMT on a CR testset using both human and automatic evaluation, and find that pretrained knowledge can indeed assist the downstream NMT model in making commonsensible predictions. Examples of the translation are provided in Table 1.

Evaluation (§4&5) Based on the strong testbed PT-NMT, we introduce how to conduct a more rigorous evaluation of CR in NMT, which is the prerequisite for conducting related research such as improving CR ability of NMT. We discuss the limitation of the existing evaluation method (He et al., 2020), and reveal the necessity of considering NMT candidates in evaluating CR ability of NMT. Furthermore, we propose a novel entity-aware automatic evaluation method, which takes into account the importance of certain words in the translation candidates that require commonsense knowledge to be translated accurately.

Challenge (§6) Our findings indicate that the arbitrary integration of extra knowledge, such as forward-translation (FT; Zhang and Zong, 2016) and back-translation (BT; Sennrich et al., 2016) does not always lead to an improvement in CR ability of NMT models and may even introduce negative effects. To address this challenge, we suggest potential research directions, including the enhancement of NMT encoder and the better utilization of target monolingual data. We also conduct a preliminary experiment to validate this hypothesis and hope that our methods and findings will provide new insights into the field of CR in NMT and inspire further advancements in this area.

Our **main contributions** are as follows:

- We demonstrate the effectiveness of incorporating pretrained knowledge into NMT models, and establish these models as robust testbeds for investigating CR in NMT.
- We reveal the limitation of the existing evaluation method for CR in NMT, and propose the use of candidate-aware metrics as a more effective and reliable alternative.
- We propose a novel entity-aware evaluation method, which is more aligned with human judgment and provides a more reliable evaluation of CR ability of NMT models.
- We identify challenges in improving CR ability of NMT, and suggest directions for further research, e.g., utilizing target monolingual data and enhancing the encoder module.

2 Background

Commonsense Reasoning Testset in NMT We first provide a brief overview of the CR testset investigated in He et al. (2020)[1]. Each instance of the testset is a triple (x, y^r, y^c), where x stands for a source sentence, and two English references (i.e., a right one y^r and a contrastive one y^c) are created for each source sentence with the intention to demonstrate how different interpretations of an ambiguity point would affect the translation results, and therefore forming an instance as follows:

x 学校 规定 学生 上学 要 背 书包。
y^r The school requires students to carry school bags.
y^c The school requires students to recite school bags.

where "recite" forms the ambiguous translation.

Three subsets of source sentences are created according to three main categories: contextless syntactic ambiguity (CL-SA) with 450 instances, contextual syntactic ambiguity (CT-SA) with 350 instances, and lexical ambiguity (LA) with 400 instances. For more details, refer to Appendix A.1.

Automatic Evaluation of CR in NMT The vanilla evaluation method proposed by Sennrich (2017); He et al. (2020) evaluates CR accuracy of NMT by comparing the prediction probability of a right reference y^r to that of its corresponding contrastive reference y^c. If an NMT model assigns a higher prediction score to the right reference than to the contrastive one, the model is considered to

[1] https://github.com/tjunlp-lab/CommonMT

Type	Prob		Human	
	NMT	PT-NMT	NMT	PT-NMT
CL_SA	67.1	$71.1_{+4.0}$	71.8	$74.2_{+2.4}$
CT_SA	56.3	$59.4_{+2.9}$	55.7	$62.3_{+6.6}$
LA	61.5	$63.3_{+1.8}$	62.5	$65.5_{+3.0}$
ALL	62.1	$65.1_{+3.0}$	64.0	$67.8_{+3.8}$

Table 2: CR accuracy measured by human evaluators and automatic evaluation metrics. PT-NMT gets consistently higher CR accuracy than NMT.

have made a commonsensible prediction. The final CR accuracy is calculated over the whole testset:

$$\text{ACC}_{\text{PROB}} = \frac{1}{I}\sum_{i=1}^{I} \mathbb{1}_{P_{\text{NMT}}(y_i^r|x_i) > P_{\text{NMT}}(y_i^c|x_i)} \quad (1)$$

where I denotes the number of instances in the testset. We name this evaluation as PROB in the following part. PROB is a widely-used metric to evaluate contrastive evaluation of sequence-to-sequence learning tasks (Vamvas and Sennrich, 2021a,b).

3 Commonsense Reasoning in PT-NMT

This section aims to answer the question of whether the incorporation of pretrained knowledge can improve CR ability of NMT models.

3.1 Setup

Experimental Data To make a fair comparison, we follow He et al. (2020) to use the CWMT Chinese-English corpus as the training set (about 9M)[2]. The validation set is newstest2019 and the in-domain testset is newstest2020. We use the CR testset mentioned in He et al. (2020) to evaluate CR ability of NMT, and compare the performance of existing automatic evaluation metrics. We use the mBART tokenizer (Liu et al., 2020b) to directly tokenize the raw text and split the text into sub-words for both Chinese and English.

Translation Models We mainly compare two model types: vanilla and pretraining-based. For NMT, we train it using the setting of the scale Transformer (Ott et al., 2018) with large-batch training of nearly 460K tokens per batch. This setting of using large batch size helps to enhance model training. One notable setting is that the dropouts for hidden states/attention/relu are set to 0.3/0.1/0.1, and the training step is 50K. For PT-NMT, we use the pretrained sequence-to-sequence model mBART (Liu et al., 2020b)[3] as our testing ground due to its high reliability and reproducibility (Tang et al., 2021; Liu et al., 2021b). All the settings follow the mBART paper, except we use a batch size of 32K and fine-tune the mBART model on the CWMT corpus for 100K steps. The training process of mBART takes more steps than that of the vanilla transformer. The reason is that this process can be seen as a fine-tuning process of a large language model. A small learning rate is necessary to achieve optimal learning performance, which in turn makes the overall training process longer. For both models, we select the checkpoint with the lowest validation perplexity for model testing. The beam size is 4 and the length ratio is 1.0.

3.2 Results

To start with, we compare the in-domain translation performance of the NMT and PT-NMT models. The two models achieve comparable BLEU scores of 25.9 and 26.2, respectively. These results are in line with previous studies (Liu et al., 2020b), which have shown that pretrained knowledge does not lead to significant improvements in high-resource settings. However, as our following human and automatic evaluations will show, there is a noticeable difference in CR abilities of the two models.

Human Evaluation To evaluate the impact of pretrained knowledge on CR ability of NMT models, we first conduct a human evaluation of the NMT candidates of NMT and PT-NMT. The evaluation involves two bilingual experts who are asked to label whether an NMT candidate is commonsensible, with the assistance of the right and contrastive references in the testset. In case of conflicting labels provided by the two experts, they engage in a discussion to arrive at a final decision on the appropriate label to be assigned.

PT-NMT Is Better in CR Table 2 illustrates CR accuracy of NMT and PT-NMT measured by human and automatic evaluation. The results indicate that PT-NMT achieves substantial enhancements in CR compared to NMT across all the subsets. This suggests that the knowledge obtained from large-scale pretraining assists the downstream NMT model in making commonsensible predictions.

[2] http://nlp.nju.edu.cn/cwmt-wmt

[3] https://github.com/pytorch/fairseq/blob/main/examples/mbart/README.md

Type	Metric	NMT			PT-NMT		
		χ^2	τ	α	χ^2	τ	α
CL_SA	**BLEU**	107.7	0.430	124.8	95.9	0.413	123.2
	PROB	115.6	0.419	138.1	79.5	0.391	121.9
	BLEURT	115.6	0.474	217.3	115.6	0.431	161.2
	BERTS.	**154.1**	**0.507**	**252.5**	**158.6**	**0.486**	**225.9**
CT_SA	**BLEU**	84.1	0.450	95.9	84.3	0.468	125.7
	PROB	80.1	0.401	84.5	56.4	0.372	66.7
	BLEURT	**124.9**	0.498	134.6	101.7	0.466	111.1
	BERTS.	113.8	**0.501**	**178.7**	**129.8**	**0.500**	**154.0**
LA	**BLEU**	90.8	0.484	129.7	86.0	0.496	124.5
	PROB	152.9	0.502	189.6	156.2	0.485	173.0
	BLEURT	**204.0**	**0.568**	**328.4**	**182.5**	**0.571**	**347.0**
	BERTS.	176.7	0.535	278.5	159.2	159.2	257.4
ALL	**BLEU**	288.5	0.458	315.7	274.5	0.451	329.4
	PROB	351.3	0.455	414.0	289.4	0.429	359.7
	BLEURT	445.3	0.519	608.6	396.5	0.493	549.8
	BERTS.	**448.5**	**0.525**	**691.9**	**449.8**	**0.502**	**613.7**

Table 3: Meta-evaluation of the sentence-level metrics. χ^2, τ and α represents the chi-square test, Kendall's τ and ANOVA, respectively. BERTS. denotes BERTScore. All the p-values obtained are < 0.01.

4 Improving Automatic Evaluation of Commonsense Reasoning

Based on the strong testbed of PT-NMT, in this section, we re-examine the existing automatic evaluation methods, and further enhance the evaluation.

4.1 Candidate-Aware Metrics

Limitation of PROB We begin by discussing the limitations of the existing metric PROB. We argue that PROB is a suboptimal metric for evaluating CR in NMT, as it ignores the most important and direct aspect of NMT: the candidates. The fact that an NMT model gives a high prediction score to the right reference does not guarantee that it will produce a commonsensible candidate, due to the bias or errors of the NMT search algorithm (Stahlberg and Byrne, 2019). A more suitable approach for evaluating CR would be to consider the NMT candidate y' as part of the evaluation process, to align it more closely with human judgement.

CR Accuracy Calculation To achieve this goal, we propose to evaluate CR prediction by directly comparing the similarities between a candidate and a pair of right and contrastive references. If the NMT candidate is more similar to the right reference than the corresponding contrastive one (i.e., $\text{sim}(y^r, y') > \text{sim}(y^c, y')$), the NMT model is considered to have made a correct prediction.

We choose three representative automatic metrics (i.e., the $\text{sim}(\cdot)$ function) for calculating the similarity: the most widely-used BLEU (Papineni et al., 2002) and the two powerful PT-based metrics BLEURT (Sellam et al., 2020) and BERTSCORE (Zhang et al., 2020). The final accuracy is a statistic of the whole testset. For example, the CR accuracy of BLEU is:

$$\text{ACC}_{\text{BLEU}} = \frac{1}{I}\sum_{i=1}^{I} \mathbb{1}_{\text{BLEU}(y_i^r, y_i') > \text{BLEU}(y_i^c, y_i')} \quad (2)$$

where I denotes the number of instances in the testset. Similar equations are used for BLEURT and BERTSCORE. We believe that these metrics can better reflect CR ability of an NMT model as they take into account the NMT candidates, which is a critical aspect of NMT. Appendix A.2 gives CR accuracy of each metric in NMT and PT-NMT.

4.2 Meta-Evaluation

Settings The above human evaluation enables the meta-evaluation of the metric performance in evaluating CR ability. We conduct chi-square tests, analysis of variance (ANOVA) and calculate Kendall rank correlation coefficients (Kendall's τ) between labels given by human evaluators and evaluation results of each metric. (1) In the chi-square test, we aim to determine the presence of a significant association between labels assigned by human evaluators and those predicted by our metrics. We use a binary classification approach, by comparing the scores of the right references against those of the contrastive references. Examples with a higher score on the right side are classified as positive, while those with an equal or lower score on the right side are classified as negative. We then construct contingency tables using the human labels and the predicted labels and conduct the chi-square test on these tables to determine the significance of the association between the two sets of labels. (2) In the ANOVA and Kendall's τ, we treat the difference in scores between the two sides as a continuous feature and the human labels as a categorical variable, where positive is represented as 1 and negative as 0. The ANOVA and Kendall's τ tests aim to determine if there is a strong correlation between the feature and the category. For Kendall's τ, we calculate the τ_b statistic which makes adjustments for tied pairs in the data. All the test results are reported in a way that a higher value indicates a stronger correlation to human judgement.

BERTSCORE Wins The results are shown in Table 3. BLEU underperforms the other metrics since

x	当/When 地震/earthquake 袭击/hit 日本/Japan 时，援助的/assisting 是/is 中国/China。
y^r	When the earthquake hit Japan, China has assisted.
e^r	{assisted}
y^c	When the earthquake hit Japan, China was aided.
e^c	{aided}

Table 4: Examples of commonsense entities in the testset. Highlights denote the ambiguous points, of which the meanings decide the correctness of the translations.

its design principle is at the corpus-level instead of the sentence-level, besides it fails to handle semantic and syntactic variants. Encouragingly, the two PT-based candidate-aware metrics BLEURT and BERTSCORE consistently achieve better correlations than the widely-used metric PROB. The abundant commonsense knowledge embedded in pretrained language models helps them to judge correctly. This observation confirms our assumption that CR evaluation can benefit from being aware of the NMT candidates. Overall, the results of the ALL testset indicate that BERTSCORE achieves superior performance in comparison to BLEURT in terms of correlation. However, the undesired performance of BERTSCORE in the LA testset motivates us to further investigate the automatic evaluation metrics for CR in NMT.

5 Entity-Aware BERTSCORE

In this section, we will further investigate BERTSCORE and propose a novel method for enhancing its correlation to human judgement by introducing the commonsense entity in CR of NMT.

5.1 Method

Commonsense Entity Upon examination of the instances in the CR testset, as depicted in Table 4, it is evident that the majority of the differences between the right and contrastive references are minor. These variations often pertain to the ambiguous elements in the source sentence, which play a crucial role in determining the commonsense nature of a translation generated by an NMT model. To enhance the correlation of BERTSCORE with human judgement, we propose to increase the weight of these elements during the calculation of the metric by leveraging their significance in the evaluation of commonsense in translations. We define these elements as commonsense entities.[4] The sets of commonsense entities in the right and contrastive references are defined as follows:

$$e^r = \{t | t \in y^r \wedge t \notin y^c\} \quad (3)$$

$$e^c = \{t | t \notin y^r \wedge t \in y^c\} \quad (4)$$

where y^r and y^c indicate the right and contrastive references of the source sentence x_i, respectively. Specifically, the commonsense entities in the right set are the words that only appear in the right reference. Similarly, the contrastive set contains words that only appear in the contrastive reference.

Integrating Weight into BERTScore We first briefly introduce the original BERTScore:

$$S_{\text{BERT}}(y, y') = \frac{1}{|y|} \sum_{t \in y} \max_{t' \in y'} d(t, t') \quad (5)$$

where y and y' represent the reference and NMT candidate, respectively, and $d(t, t')$ denotes pairwise similarity between word embeddings of t and t'. In this original method, every word in the reference sentence is given equal weight while calculating the average similarity score, without considering the crucial words related to ambiguous points (i.e., commonsense entities).

To address this limitation, we propose the entity-aware BERTSCORE. For the score between the right reference and NMT candidate, the right commonsense entities are assigned a greater weight:

$$S_{\text{EntBERT}}(y^r, y') = \frac{\sum_{t \in y^r} m(t) \max_{t' \in y'} d(t, t')}{\sum_{t \in y^r} m(t)} \quad (6)$$

where $m(t)$ represents the score weight for word t. If $t \in e^r$, $m(t)$ is set to a value greater than 1, otherwise it is set to 1. This approach ensures that candidates that can accurately translate commonsense entities will receive a higher score on the right side. Similarly, for the score calculated on the contrastive reference, if $t \in e^c$, the weight $m(t)$ is also set to a value greater than 1. Candidates that translate commonsense entities incorrectly will also receive a higher score on the contrastive side.

Calculation Filtering In certain instances of the testset, the right and contrastive reference may have different syntactic structures or wording, even though they convey similar meanings. This issue can lead to a large number of words in both commonsense entity sets, many of which may not be directly related to the ambiguous point in the current instance. To address this, we propose to only

[4] We exclude stopwords and punctuation from our definition of commonsense entities.

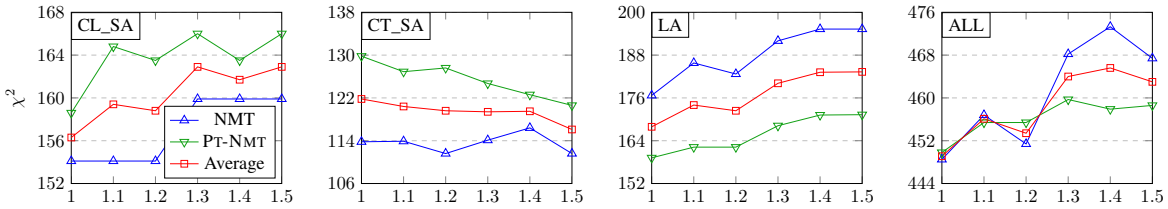

Figure 1: Chi-square test results χ^2 between labels given by human evaluators and predicted by BERTSCORE when commonsense entities are assigned different weights. The original BERTScore can be seen as setting $\lambda = 1$ that the weights of all words are treated equally.

apply commonsense entity weights on those sentences whose commonsense entity set contains no more than 3 words. This approach helps filter out sentences for which our proposed metric may not be suitable and also avoids assigning unnecessary weight to unrelated words.

5.2 Correlations with Human Judgments

Settings To determine the optimal weight for commonsense entities, we conduct experiments on the testset, varying the weight from 1 to 1.5. For each weight, we calculate the corresponding BERTSCORE between the candidates and references and then determine the predicted labels by comparing the scores for the right and contrastive references. We then perform a chi-square test on these predicted labels human labels. The baseline for this experiment is a commonsense entity weight of 1.0, where all words in the candidates are treated equally without considering their importance.

Results The results are shown in Figure 1. It is observed that by increasing the commonsense entity weight from 1.0, there is an overall improvement in the performance of evaluating CR abilities of NMT and PT-NMT. The performance reaches its peak when the weight is around 1.4. As a result, 1.4 is set as the weight for commonsense entities, and 1 is for other trivial words. The entity-aware BERTSCORE on the contrastive side is calculated in a similar manner, only replacing the right commonsense entity set e^r with the contrastive commonsense entity set e^c. The results validate the effectiveness of our method. Based on these results, in the following part, we mainly use the entity-aware BERTScore with $\lambda = 1.4$ as the default automatic evaluation metric for CR in NMT.

6 Challenge in Commonsense Reasoning

The above results sufficiently validate the positive impact of CR ability brought by pretraining. In this section, we prob the use of unlabeled data to improve NMT performance, with a focus on determining their impact on CR ability and identifying potential areas for improvement.

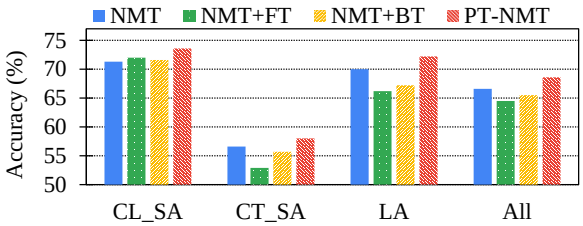

Figure 2: CR accuracy of combining NMT with FT and BT. Both variants underperform NMT and PT-NMT.

6.1 Probing of Monolingual Data Utilization

This experiment aims to study the research question: *How do the additionally learned monolingual (unlabeled) data impact CR ability?* We compare the results of mainstream data augmentation methods: forward-translation (FT) and back-translation (BT), making use of source and target monolingual data respectively. In general, the process of incorporating monolingual data for NMT training involves translating source monolingual data into the target language (i.e., FT) and back-translating target monolingual data (i.e., BT). This synthetic data is then combined with the original bilingual data for training the NMT model.

Setup and In-domain BLEU All the synthetic data is generated by vanilla NMT. We do not utilize PT-NMT to avoid mitigating any potential biases that may be introduced from pretrained knowledge. The FT data is generated by NMT using the samples from the combination of WMT19 and WMT20 Chinese News monolingual data. The BT data is generated by a reversed NMT using the samples from WMT16 English News monolingual data. The sampling ratio to the original training data is 1:1. The text preprocessing and model training keep

Testset	NMT			PT-NMT				
	Vanilla	+FT	+BT	Vanilla	+FT	+BT	+TagBT	+LAttn
In-domain	25.9	26.7	26.4	26.2	26.8	26.8	26.8	26.5

Table 5: BLEU scores of the trained NMT models on the in-domain newstest2020 testset. Combining NMT and PT-NMT with the other methods can gain in-domain model improvements.

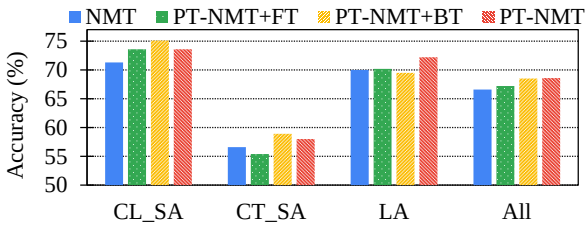

Figure 3: CR accuracy of combining PT-NMT with FT and BT. Combining PT-NMT with BT improves the accuracy of syntactic ambiguity but decreases the accuracy of lexical ambiguity.

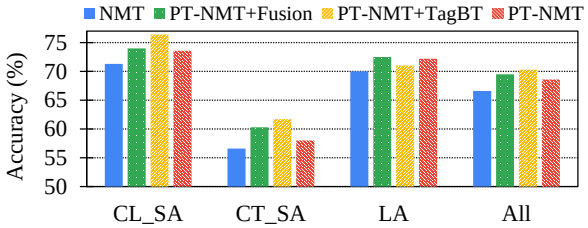

Figure 4: CR accuracy of preliminary validation. Both enhancements of the encoder module (+Fusion) and target monolingual data utilization (+TagBT) can improve CR accuracy of PT-NMT.

unchanged. The in-domain BLEU scores of the models in the following sections are shown in Table 5. Overall, combining NMT and PT-NMT with the other data augmentation methods can gain in-domain improvement, but their CR abilities show large differences, further accentuating the necessity of broadening the understanding.

PT-NMT vs. FT and BT Figure 2 shows that solely using FT and BT underperform NMT and PT-NMT. One possible reason is the unexpected noises/errors during the generation of synthetic data, hindering the model from learning commonsense knowledge. Differently, pretraining utilizes monolingual data in an end-to-end manner, alleviating the risk of error propagation and thus bringing more benefits to the CR ability. Pretraining is superior to other data augmentation methods in enhancing CR ability of NMT.

PT-NMT with FT and BT PT-NMT has learned abundant semantic and syntactic knowledge during pretraining, which is competent to learn synthetic data. This experiment investigates the effect of combing PT-NMT with FT and BT, as shown in Figure 3. (1) When combining PT-NMT with FT, we observe that the model fails to demonstrate any improvement in all cases. One potential explanation for this is that FT primarily enhances the understanding of source sentences, as per previous research, whereas PT also primarily improves the understanding of source sentences, as reported in Liu et al. (2021a). This implies that these two methods may not be complementary in nature. (2) When combining PT-NMT with BT, the model gains satisfactory improvements on the two syntactic ambiguity testsets, indicating that target data contributes to overcoming such ambiguities. The pity is the decreased performance in LA, but this is reasonable since the BT process indeed hurts the lexical diversity of source data (Nguyen and Chiang, 2018), leading to noisy signals that worsen the learning of source features. Overall, combining PT-NMT with BT helps alleviate syntactic ambiguity but worsens the ability to address lexical ambiguity.

6.2 Potential Directions

This part provides a preliminary study to validate the above findings. The scope of this paper is not to exhaustively explore the entire space, but to demonstrate that the findings are convincing and can guide future studies related to CR in NMT.

Encoder Module Enhancement The encoder module can play a greater role in improving CR of NMT. PT mainly improves the encoder module of NMT and has shown consistent improvement. Therefore, if we continue to leverage the encoder module, for example, by augmenting it with other types of knowledge or further enhancing its encoding ability, CR ability of NMT should also be improved. This is supported by previous studies in the area of word sense disambiguation (Tang et al., 2019), which also highlights the importance of enhancing the encoder module (Liu et al., 2021c).

Preliminary Validation: we further evaluate the effectiveness of a simple and effective encoder layer fusion method (Bapna et al., 2018) that aims to improve the learning of source representations. Specifically, this method differs from the vanilla NMT approach, which connects each decoder layer only to the topmost encoder layer, by allowing each decoder layer to extract features from all encoder layers, including the encoder embedding layer. This strengthens the model ability to learn and utilize encoder features. The results in Figure 4 show that this simple method can improve the overall CR ability, which confirms the effectiveness of this method in improving the encoder. It is likely that incorporating additional useful knowledge may yield even more significant benefits (Wang et al., 2022b; Li et al., 2022).

Utilization of Target Monolingual Data Previous sections have shown that combing PT-NMT with BT brings both positive and negative impacts on the CR ability, and one of the possible reasons for the negatives is that the source-side BT data contains too many noises and lower lexical richness. Since BT is a popular line of research in NMT, digging it more in the future might bring an efficient improvement of CR in NMT.

Preliminary Validation: we improve BT by adding a tag to the BT data (TagBT; Caswell et al., 2019) to make the model can selectively learn more syntactic knowledge and less lexical knowledge from the BT data. Figure 4 shows that the result of this simple method meets our expectation that the CR ability of each type has been strengthened, especially the ability to solve lexical ambiguity.

7 Related Work

7.1 Pretraining in NMT

Pretraining learns abundant syntactic and semantic knowledge from large-scale unlabeled data, which has been sufficiently validated to be useful for various downstream tasks (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019). This kind of knowledge can also boost the performance of NMT, especially for those translation directions whose labeled data (i.e., parallel corpus) are scarce.

The first research line investigates how to better model the interdependency between knowledge embedded in pretrained models and NMT models, such as introducing downstream task-aware: 1) pretraining architectures (Song et al., 2019; Conneau et al., 2020; Lewis et al., 2020b,a); 2) pretraining strategies (Liu et al., 2020b; Yang et al., 2020b; Lin et al., 2020; Ren et al., 2021; Sadeq et al., 2022; Wang et al., 2022a); and 3) knowledge extractors (Yang et al., 2020a; Zhu et al., 2020). These studies continuously strengthen the useful pretrained knowledge for NMT models.

Previous studies have attempted to understand the improvements in NMT models resulting from pre-training. Understandings at the model level investigate the contribution of different NMT modules to the improvement (Rothe et al., 2020; Cooper Stickland et al., 2021; Gheini et al., 2021). Insights at the data level compare pretraining with BT and find that pretraining is a nice complement to BT (Liu et al., 2021a; Huang et al., 2021; Deng et al., 2022). Liu et al. (2021b) explore the copying ability of vanilla NMT and PT-NMT, and reveal the importance of understanding model ability. Our work builds on this research by providing a systematic examination of CR ability of NMT and proposing potential directions for further enhancement.

7.2 Commonsense Reasoning in NMT

CR is an important task for NLP, whose design principle is investigating whether a model goes beyond pattern recognition or not. Translation models equipped with commonsense are expected to better deal with word sense disambiguation (WSD), complex linguistic structures, and other challenging translation tasks (Bar-Hillel, 1960). WSD is a major source of translation errors in NMT, and the solution of which relies heavily on the model ability of context understanding or CR. Rios Gonzales et al. (2017) introduce a testset for WSD, and Rios et al. (2018) enhance the testset and propose a novel semi-automatic human evaluation. Tang et al. (2019) find that the NMT encoder is highly relevant to the ability of WSD. Emelin et al. (2020) attribute the unsatisfactory WSD performance to not only model learning but also the data bias of the training set. Additionally, pronoun resolution in sentences with complex linguistic structures is another task of CR (Levesque et al., 2012). Davis (2016) introduce how to evaluate pronoun resolution in MT and explains the difficulty in its evaluation.

While previous works mainly focus on studying one specific type of CR, He et al. (2020) introduce a customized benchmark covering the above types to directly evaluate CR ability of NMT. Based on this benchmark, Huang et al. (2021) observe the

CR ability enhancement by utlizing PT knowledge. However, the internal cause of the improvement is still unclear and the evaluation part of CR in NMT can be further improved. In the presenting paper, we provide rational explanations for the improvement based on evaluation and probing methods, which can inform the development of future translation systems with stronger CR capabilities.

8 Conculsion

This paper expands on the understanding of commonsense reasoning in NMT. We confirm the superior commonsense reasoning ability of pre-training enhanced NMT models through both automatic and human evaluations. We introduce a novel entity-aware evaluation metric that takes into account the NMT candidates to alleviate the limitations of existing metrics. Based on the enhanced evaluation metric, we identify the challenges and potential research directions for further enhancing the commonsense reasoning ability of NMT, including the further enhancement of the encoder module and utilization of target monolingual data.

Limitations

Research on commonsense requires a good understanding of bilingual knowledge. While this article focuses on evaluating commonsense reasoning ability of vanilla NMT and pretraining-based NMT models on a Chinese-English testset, testing commonsense reasoning ability on more language pairs would provide a more comprehensive understanding of the commonsense reasoning ability of NMT models, to further enhance the research of commonsense reasoning in NMT.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62206076), the Science and Technology Development Fund, Macau SAR (Grant Nos. FDCT/060/2022/AFJ, FDCT/0070/2022/AMJ), Shenzhen College Stability Support Plan (Grant Nos. GXWD20220811173340003, GXWD20220817123150002), Shenzhen Science and Technology Program (Grant No. RCBS20221008093121053) and the Multi-year Research Grant from the University of Macau (Grant No. MYRG2020-00054-FST). We would like to thank the anonymous reviewers and meta-reviewer for their insightful suggestions.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ankur Bapna, Mia Chen, Orhan Firat, Yuan Cao, and Yonghui Wu. 2018. Training deeper neural machine translation models with transparent attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3028–3033, Brussels, Belgium. Association for Computational Linguistics.

Yehoshua Bar-Hillel. 1960. The present status of automatic translation of languages. *Advances in computers*, 1:91–163.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. 2021. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3440–3453, Online. Association for Computational Linguistics.

Ernest Davis. 2016. Winograd schemas and machine translation. *ArXiv preprint*, abs/1608.01884.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103.

Hexuan Deng, Liang Ding, Xuebo Liu, Meishan Zhang, Dacheng Tao, and Min Zhang. 2022. Improving simultaneous machine translation with monolingual data. *arXiv preprint arXiv:2212.01188*.

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.

Denis Emelin, Ivan Titov, and Rico Sennrich. 2020. Detecting word sense disambiguation biases in machine translation for model-agnostic adversarial attacks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7635–7653, Online. Association for Computational Linguistics.

Zhiwei Feng, State Language Commission, et al. 1995. On potential nature-of ambiguous construction. *Journal of Chinese Information Processing*, 4:14–24.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Mozhdeh Gheini, Xiang Ren, and Jonathan May. 2021. Cross-attention is all you need: Adapting pretrained Transformers for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1765, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jie He, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672, Online. Association for Computational Linguistics.

Dandan Huang, Kun Wang, and Yue Zhang. 2021. A comparison between pre-training and large-scale back-translation for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1718–1732, Online. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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

Zhaocong Li, Xuebo Liu, Derek F. Wong, Lidia S. Chao, and Min Zhang. 2022. ConsistTL: Modeling consistency in transfer learning for low-resource neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8383–8394, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020a. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436, Online. Association for Computational Linguistics.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021a. On the complementarity between pre-training and back-translation for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2900–2907, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2021b. On the copying behaviors of pre-training for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4265–4275, Online. Association for Computational Linguistics.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, and Zhaopeng Tu. 2021c. Understanding and improving encoder layer fusion in sequence-to-sequence learning. In *International Conference on Learning Representations*.

Xuebo Liu, Derek F. Wong, Yang Liu, Lidia S. Chao, Tong Xiao, and Jingbo Zhu. 2019. Shared-private bilingual word embeddings for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3613–3622, Florence, Italy. Association for Computational Linguistics.

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

Toan Nguyen and David Chiang. 2018. Improving lexical choice 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 334–343, New Orleans,

Louisiana. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

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.

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.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, OpenAI.

Shuo Ren, Long Zhou, Shujie Liu, Furu Wei, Ming Zhou, and Shuai Ma. 2021. SemFace: Pre-training encoder and decoder with a semantic interface for 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 4518–4527, Online. Association for Computational Linguistics.

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The word sense disambiguation test suite at WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 588–596, Belgium, Brussels. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Nafis Sadeq, Canwen Xu, and Julian McAuley. 2022. Informask: Unsupervised informative masking for language model pretraining. *ArXiv*, abs/2210.11771.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

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

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? 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 3356–3362, Hong Kong, China. Association for Computational Linguistics.

Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pre-training is (almost) all you need: An application to commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3878–3887, Online. Association for Computational Linguistics.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2019. Encoders help you disambiguate word senses in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1429–1435, Hong Kong, China. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association*

for Computational Linguistics: ACL-IJCNLP 2021, pages 3450–3466, Online. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2021a. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2021b. On the limits of minimal pairs in contrastive evaluation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 58–68, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wenxuan Wang, Wenxiang Jiao, Yongchang Hao, Xing Wang, Shuming Shi, Zhaopeng Tu, and Michael Lyu. 2022a. Understanding and improving sequence-to-sequence pretraining for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2591–2600, Dublin, Ireland. Association for Computational Linguistics.

Zhijun Wang, Xuebo Liu, and Min Zhang. 2022b. Breaking the representation bottleneck of Chinese characters: Neural machine translation with stroke sequence modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6473–6484, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020b. CSP:code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636, Online. Association for Computational Linguistics.

Hui Yuan. 2001. *A Contemporary Chinese Polysemy Dictionary*, 2. edition. Shu Hai.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9733–9740.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Contextless Syntactic Ambiguity (CL-SA; 450 Instances)	
Source	捕获的/Hunting 是/is 猎人/hunter 。
Right	The hunter is hunting .
Contrast	The hunter is hunted .
Contextual Syntactic Ambiguity (CT-SA; 350 Instances)	
Source	手术/surgery 开刀的/operated 是/is 他父亲/his father，因为/because 他父亲/his father 得了重病/get serious illness。
Right	It was his father who underwent surgery, because his father was seriously ill.
Contrast	It was his father operated the surgery, because his father was seriously ill.
Lexical Ambiguity (LA; 400 Instances)	
Source	学校/school 规定/mandates 学生/students 上学/go to school 要/must 背/carry 书包/school bag。
Right	The school requires students to carry school bags.
Contrast	The school requires students to recite school bags.

Table 6: Examples of the three kinds of ambiguities on the commonsense reasoning testset. Highlights denotes the ambiguity part requiring commonsense knowledge for accurate translation.

Type	NMT				
	Human	**BLEU**	**PROB**	**BLEURT**	**BERTS.**
CL_SA	71.8	59.6	67.1	67.1	71.3
CT_SA	55.7	49.4	56.3	53.4	57.4
LA	62.5	48.8	61.5	65.8	69.5
ALL	64.0	53.0	62.1	62.7	66.7
	PT-NMT				
	Human	**BLEU**	**PROB**	**BLEURT**	**BERTS.**
CL_SA	74.2	63.6	71.1	67.8	73.1
CT_SA	62.3	49.7	59.4	57.7	58.3
LA	65.5	51.5	63.3	67.5	72.3
ALL	67.8	55.5	65.1	64.8	68.5

Table 7: Commonsense reasoning accuracy measured by different metrics. PT-NMT gets higher commonsense reasoning accuracy than NMT.

A Appendix

A.1 Commonsense Reasoning Testset

He et al. (2020) construct a testset for testing the commonsense reasoning in automatic Chinese⇒English translation task, and the three rules by which the testset is constructed are as follows: 1) Target candidates are intended to be tested since the commonsense knowledge of NMT will be in this case more identifiable; 2) The testset covers three types of ambiguity point, which are contextless syntactic ambiguity (CL_SA), contextual syntactic ambiguity (CT_SA), and lexical ambiguity (LA); 3) Two English translations are created for each source sentence with the intention to demonstrate how different interpretations of the ambiguity point would affect the translation results. The construction of these test instances is based on the true cases that might cause ambiguities in Chinese and English. The polysemous words contained in the LA set are chosen from a Chinese polysemy dictionary (Yuan, 2001). As for the CT_SA and CL_SA subsets, the test instances are based on adopted 12 Chinese structures (Feng et al., 1995) that may result in ambiguities in English.

Table 6 shows some examples of the testset. Each English translation that correctly rendered the Chinese source sentence is combined with a contrastive (incorrect) translation to form a test sample. Specifically, the first source sentence, which is retrieved from the testset of CL_SA, requires the NMT to possess commonsense knowledge about the relation between "hunter" and "prey". In the CT_SA, to understand the semantic relation between "operation" and "his father", the model has to know the commonsense knowledge that an ill person needs surgery from the sentence context. And in the last LA set, the source sentence and the translations are constructed on the basis of different interpretations of the polysemy like "背 (to recite/to carry on one's back)", where the model needs commonsense knowledge to know "school bag" is used for carrying.

A.2 Commonsense Reasoning Accuracy

In addition to the human evaluation and model probability evaluation shown in Table 2, more metrics of evaluating commonsense reasoning accuracy and their corresponding results can be found in Table 7. It can be seen that the pretrained NMT model (PT-NMT) outperforms the vanilla NMT model across all given metrics and subsets in terms of commonsense reasoning. This suggests that PT-NMT demonstrates a significant improvement in commonsense reasoning ability compared to NMT.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

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

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

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

B ☑ Did you use or create scientific artifacts?

Section 3.1

☑ B1. Did you cite the creators of artifacts you used?
Section 3.1

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
All the data and code used in this work are open-sourced.

☐ 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.?
Not applicable. Left blank.

☑ 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.
Section 3.1

C ☑ Did you run computational experiments?

Section 3&4&5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3.1&4.2&5.2

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?
Section 3.1&4.2&5.2

☑ 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?
Section 3.2&4.2&5.2

☑ 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.)?
Section 3.1&4.1&4.2

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

Section 3.2

☑ 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.?
Section 3.2

☐ 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)?
Not applicable. Our human annotation process required the labeling of only 1,000 instances and demanded a high degree of attention to detail and expertise in bilingualism. As a result, the task was completed solely by members of our research group.

☐ 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. The data used in this study is available for use in research purposes.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. We do not collect data.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3.2