GCPG: A General Framework for Controllable Paraphrase Generation

Kexin Yang♠ * Dayiheng Liu♠ † Wenqiang Lei♢ Baosong Yang♠ Haibo Zhang♠
Xue Zhao♠ Wenqing Yao♠ Boxing Chen♠
♠Alibaba Group
♢National University of Singapore
{kexinyang0528, losinuris}@gmail.com

Abstract

Controllable paraphrase generation (CPG) incorporates various external conditions to obtain desirable paraphrases. However, existing works only highlight a special condition under two indispensable aspects of CPG (i.e., lexically and syntactically CPG) individually, lacking a unified circumstance to explore and analyze their effectiveness. In this paper, we propose a general controllable paraphrase generation framework (**GCPG**), which represents both lexical and syntactical conditions as text sequences and uniformly processes them in an encoder-decoder paradigm. Under GCPG, we reconstruct commonly adopted lexical condition (i.e., *Keywords*) and syntactical conditions (i.e., *Part-Of-Speech sequence*, *Constituent Tree*, *Masked Template* and *Sentential Exemplar*) and study the combination of the two types. In particular, for *Sentential Exemplar* condition, we propose a novel exemplar construction method — Syntax-Similarity based Exemplar (**SSE**). SSE retrieves a syntactically similar but lexically different sentence as the exemplar for each target sentence, avoiding exemplar-side words copying problem. Extensive experiments demonstrate that GCPG with SSE achieves state-of-the-art performance on two popular benchmarks. In addition, the combination of lexical and syntactical conditions shows the significant controllable ability of paraphrase generation, and these empirical results could provide novel insight to user-oriented paraphrasing.

1 Introduction

Paraphrase generation (Madnani and Dorr, 2010) refers to restating a given sentence into an alternative surface form while keeping the semantics unchanged.[1] It is of long-standing interest (McKeown, 1983), with various applications such as ques-

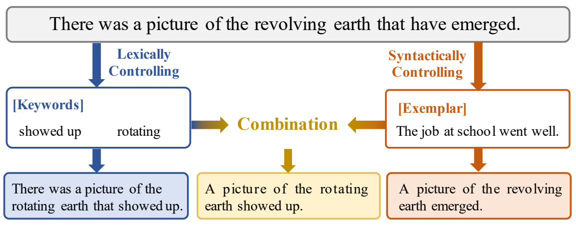

There was a picture of the revolving earth that have emerged.
Lexically Controlling
Syntactically Controlling
[Keywords]
showed up
rotating
Combination
[Exemplar]
The job at school went well.
There was a picture of the rotating earth that showed up.
A picture of the rotating earth showed up.
A picture of the revolving earth emerged.

Figure 1: A toy example to explain what effect lexically controlling and syntactically controlling have on paraphrasing.

tion answering (Gan and Ng, 2019), machine translation (Mallinson et al., 2017), and sentence simplification (Martin et al., 2020). However, a sentence can be re-expressed in various surface forms. Lacking control might result in undesirable results (Gu et al., 2019). For example, a sentence that contains an intricate syntactic structure may cause difficulties for aphasic patients (Shewan, 1985). In that case, we could attempt to paraphrase it based on syntactical control.

To obtain desirable surface forms, most recent works focus on controllable paraphrase generation (CPG) by incorporating external conditions. Existing efforts to CPG can be roughly divided into two types: lexically and syntactically CPG. Lexically CPG is concerned with *what to say*, which generates paraphrases that contain pre-specified keywords. As shown in Figure 1, a lexically CPG model needs to generate a paraphrase that contains the given keyword "showed up". To achieve it, a sequence-to-sequence model equipped with the copy mechanism is commonly used (Zeng et al., 2019). Different from lexically CPG, syntactically CPG concentrates on *how to say it*, generating a paraphrase that conforms to the syntax of a given exemplar (i.e., a sentence illustrating certain syntax patterns). Substantial efforts have been made on constructing syntactical features of the given exemplar. For example, Kumar et al. (2020) incorporate a full syntactic tree of the exemplar to

* Work is done during internship at DAMO Academy
† Corresponding author.
[1]In this paper, we make our first attempt with English corpora.

guide paraphrasing; Bui et al. (2021) construct a masked template to direct generation by masking words with certain Part-of-Speech (POS) type of exemplar; Chen et al. (2019) directly use the sentential exemplar. Since sentential exemplars are only available for testing, they have to manufacture exemplars for training by replacing certain words from the target sentence.[2] Despite the progress on the two types of conditions individually, *what to say* and *how to say it* are both important for CPG (Kumar et al., 2020). Furthermore, there is no unified framework to study the effectiveness of these conditions and their joint utilization.

To fill this gap, we propose a **G**eneral **C**ontrollable **P**araphrase **G**eneration framework (GCPG) to jointly include both lexically and syntactically CPG in a unified model. The key idea is to reconstruct both lexical and syntactical conditions as text sequences and process them in a text-to-text encoder-decoder paradigm. This also allows GCPG to easily utilize the strong language modeling capacity of pre-trained language models (PLMs), which have demonstrated great potential (Bui et al., 2021) yet rarely been explored under the topic of CPG. For the lexical condition, we concatenate the pre-specified keywords as a sequence while exploring different methods to pre-specify keywords from rule-based to model-based. As for syntactical conditions, we reconstruct commonly used syntactic features as sequences, such as Linearised Constituent Tree (Iyyer et al., 2018) and masked template based on word mask (Bui et al., 2021). Besides the manufactured syntax features, we hypothesize that directly using the exemplar is more effective as it can benefit from the powerful sentence modeling capability of PLMs. To construct the exemplar for training, we propose a novel exemplar construction method as **S**yntax-**S**imilarity based **E**xemplar (SSE). Specifically, we use a sentence that is syntactically similar but lexically different from the target sentence, which is retrieved in a self-constructed exemplar dictionary based on the training set. This is different from existing methods that construct exemplar through modifying target sentences (Chen et al., 2019), alleviating exemplar-side words copying problem (Bui et al., 2021) brought by Chen et al. (2019).

We examine GCPG on two popular benchmark datasets. Those discussions include not only performances of different conditions and their combinations, but also the effectiveness of GCPG instantiated by different PLMs. Experiments demonstrate that GCPG consistently shows strong performances when tested by three different methods to pre-specify keywords. For syntactical CPG, GCPG with SSE obtains 13.95/24.31/18.64 ROUGE-1/2/L and 16.38 BLEU-4 over the previous state-of-the-art (SOTA) model (Bui et al., 2021). Also, the combination of lexical and syntactical conditions show encouraging controllability of paraphrase generation in both quantitative and qualitative analysis. The main contributions are as follows:

- We propose **GCPG**, a general framework to jointly include both lexically and syntactically controllable paraphrasing. It is simple but effective, enabling flexible combinations of conditions by reconstructing them into text sequences and processing them in a text-to-text encoder-decoder paradigm. Those properties allow GCPG to easily adapt to mainstream pre-trained language models and utilize powerful language modeling capacity, which is rarely explored in CPG.

- We report a novel exemplar construction method **SSE** under the syntactical condition. It allows GCPG to directly model syntax information from natural sentences without any manufactured syntax features, while alleviating the exemplar-side words copying problem.

2 Related Work

In this section, we summarize existing works on syntactically and lexically CPG. Syntactically CPG generates a paraphrase constrained by a pre-specified sentence of a certain syntax structure namely exemplar. However, the exemplar is only available during inference, resulting in a key challenge: obtaining manual exemplars for existing paraphrasing training datasets is prohibitively expensive. To address this, some of the previous works construct syntactical features from target sentences during training, such as *POS Tagging*, *Constituent Tree*, *mask template* as illustrated in Table 1. For instance, SCPN (Iyyer et al., 2018) makes the first attempt to introduce Linearised Constituent Tree (LCT) of target sentence into paraphrasing, where LCT is predicted based on predefined parse templates. Similarly, GuiG (Li et al., 2020) proposes two models to expand a partial template LCT and generate paraphrasing, respectively. SOW-REAP (Goyal and Durrett, 2020) uses

[2]Details can be found in Section 4.

	Syntactical Condition			
Work	POS Tagging	Constituent Tree	Masked Template	Sentential Exemplar
SCPN (2018)	✓ (In Tree)	✓ (LCT Templates)	✗	✗
CGEN (2019)	✓ (In Exemplar)	✗	✗	✓ (Replace Words)
BCPG (2020c)	✗	✗	✓ (Randomly)	✗
GuiG (2020)	✗	✓ (Expanded LCT)	✗	✗
SGCP (2020)	✓ (In Tree)	✓ (Tree Structure)	✗	✗
SOW-REAP (2020)	✓ (In Tree)	✓ (Reordering)	✗	✗
AESOP (2021)	✓ (In Tree)	✓ (Sequence)	✗	✗
ParafraGPT (2021)	✓ (In Word MT)	✗	✓ (Certain POS)	✗
GCPG	✓ (POS Sequence)	✓ (LCT)	✓ (Certain POS)	✓ (SSE)

Table 1: A comparison of different conditions under syntactically CPG. LCT: Linearised Constituent Tree. The proposed framework GCPG reconstructs them as text sequences and we have experimented with all four forms.

LCT to reorder the source sentence then paraphrasing. AESOP (Sun et al., 2021) selects target LCT adaptively while paraphrasing. Different from using LCT, SGCP (Kumar et al., 2020) introduces a graph encoder to encode the Constituent Tree of exemplar as the condition. *masked template* replaces several words of the exemplar with a special token to form a template as the condition. BCPG (Liu et al., 2020c) follows BERT (Devlin et al., 2019) to randomly mask exemplar words, ParafraGPT (Bui et al., 2021) further masks exemplar words with certain POS types. However, Chen et al. (2019) advocate to directly utilize the sentential exemplar (the sentence) as the condition, because they believe "any syntactically valid sentence is a valid exemplar". Since exemplar is only available in testing, they construct exemplar by replacing words of the target sentence with others that have the same POS type. In addition, lexical constraints decoding is widely explored in text generation (Liu et al., 2020a, 2019; Hokamp and Liu, 2017a), such as neural machine translation (Hokamp and Liu, 2017b; Post and Vilar, 2018) and text summarizing. CTRLsum (He et al., 2020) uses target entity words as keywords to hint model while summarizing. However, lexically CPG constraints paraphrasing with pre-specified keywords, which is rarely explored but undoubtedly indispensable in CPG. Zeng et al. (2019) make the first attempt to integrate keywords with copy mechanism. Despite their progress, existing works only focus on a special condition under either lexically or syntactically CPG. In comparison, GCPG jointly includes lexically and syntactically CPG.

3 Methodology

3.1 GCPG Framework

Before introducing GCPG, we first give the definition of controllable paraphrase generation with external conditions. Given a source sentence $\boldsymbol{x}$ and a variety of conditions $\boldsymbol{c}$, the model generates paraphrase $\boldsymbol{y} = (y_1, y_2, ..., y_T)$ by:

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{c}) = \prod_{t=1}^{T} p(y_t|y_{<t}, \boldsymbol{x}, \boldsymbol{c}; \theta), \tag{1}$$

where θ are the model parameters trained by maximizing the conditional likelihood of outputs in a parallel corpus. Given this definition, the forms of conditions $\boldsymbol{c}$ might be varied, such as pre-defined keywords and Constituent Parse Tree. To uniformly encode these conditions and investigate their effectiveness, we propose a general framework GCPG. GCPG contains a standard encoder-decoder paradigm, which allows any mainstream PLMs to adapt to this task rapidly. Meanwhile, GCPG can flexibly use the combinations of included conditions by concatenating them as one sequence with "[SEP]". As shown in Figure 2, the source sentence "No one's home ?" is concatenated with optional sequential conditions by the separator signal "[SEP]", then fed into the model. Afterward, the model auto-regressively generates "Is anyone home?" as the final result.

3.2 Conditions under GCPG

3.2.1 Syntactical Condition

Syntactically CPG requests a syntax exemplar to constrain the syntax structure of paraphrase. However, exemplars are only available in the testing set

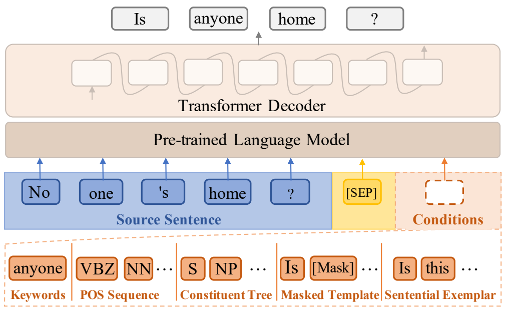


Figure 2: An overview of GCPG, the source sentence and separated condition (also being concatenated with "[SEP]") are concatenated as input.

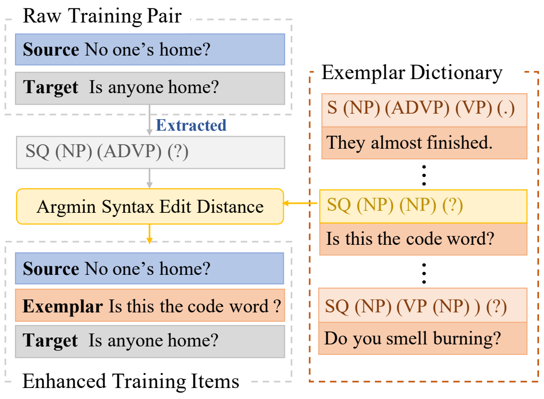


Figure 3: An overview of SSE. We take Truncated LCT as the sequential syntax structure here.

of existing paraphrasing datasets. To train a syntactically CPG model, we construct a syntactical condition based on the target sentences in the training set. During inference, we apply the same strategy to obtain the corresponding syntactical conditions from exemplars in the testing set. We explore four syntactical conditions in this work, as follows:

POS Tagging is one of simplest solutions in modeling the syntax structure (Cutting et al., 1992), which could be effectively implemented and show promising performance in various NLP tasks (Yang et al., 2021). We investigate POS Tagging as an independent condition, which is rarely explored in CPG. In detail, we extract POS sequence of target sentence by CoreNLP as the condition.[3] To learn these POS signals with PLMs, we regard these POS tokens as special ones and add them into the word vocabulary of PLMs.

Constituent Tree is a widely used condition for syntax controlling while paraphrasing. Here, we explore two kinds of LCT, i.e., full-fledged LCT and Truncated LCT. For the full-fledged LCT condition, we extract the complete sequential Constituent Tree from the target sentence for training and exemplar for testing, based on the off-the-shelf tools of CoreNLP. We further explore the Truncated LCT condition, which is the sequence that removing POS-level tokens in full-fledged LCT. Compared with full-fledged LCT, Truncated LCT drastically shortens the input length.

Masked Template is first introduced in Liu et al. (2020c), which randomly masks words of the target sentence to form a syntax template as the condition. To verify the effectiveness of this method in GCPG circumstance, we follow the current SOTA (Bui et al., 2021) to construct a masked template by substituting all nouns, adjectives, adverbs, and verbs with a special token in the exemplar. Similarly, this strategy is applied to the target sentences during training and the given exemplars during inference.

Sentential Exemplar is the most straightforward way for syntactically CPG, which directly uses the sentential exemplar as the condition. In contrast to the above three syntactical conditions, Sentential Exemplar uses natural sentences to represent desirable syntax structure, without introducing any special token which does not appear during PLMs pretraining. We argue that this way can make better use of PLMs. However, the previous method (Chen et al., 2019) suffers from the exemplar-side words copying problem during testing, which might be caused by the noticeable words overlap with the target sentence in constructing sentential exemplar during training. To alleviate this problem, we propose **S**yntax-**S**imilarity based **E**xemplar (**SSE**) to enhance sentential exemplar condition.

An overview of our SSE method is demonstrated in Figure 3. To alleviate the exemplar-side words copying issue, the proposed SSE constructs Sentential Exemplar by retrieving a syntactically similar but lexically different sentence for each target sentence during training. To achieve that, we construct an exemplar dictionary that contains the syntactical key-value mapping from the syntax structure $\boldsymbol{k}$ to its corresponding natural sentence $\boldsymbol{v}$. Each syntactical key $\boldsymbol{k} \in \boldsymbol{K}$ is a Truncated LCT sequence, and its value is a randomly selected natural sentence that can be assigned to this Truncated LCT sequence. During training, given a data pair $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ and the Truncated LCT $\boldsymbol{s}$ of $\boldsymbol{y}$, we select a syntactical key $\boldsymbol{k}^*$ by calculating the syntax edit distance D_{syn} between $\boldsymbol{s}$ and each syntactical key in the

[3]https://stanfordnlp.github.io/CoreNLP/index.html

exemplar dictionary, which can be formulated as:

$$
\begin{aligned}
\boldsymbol{k}^* &= \arg\min(D_{\text{syn}}(\boldsymbol{s}, \boldsymbol{k})) \\
&= \arg\min_{\boldsymbol{k}\in\boldsymbol{K}}\Big(\frac{\text{LevEdit}(\boldsymbol{s}, \boldsymbol{k})}{\max(|\boldsymbol{s}|, |\boldsymbol{k}|)}\Big),
\end{aligned} \tag{2}
$$

where LevEdit$(\cdot)$ denotes the token-level Levenshtein edit distance between two sequences and $|\cdot|$ denotes the token-level length of the sequence. We assign the corresponding sentence $\boldsymbol{v}^*$, which is related to $\boldsymbol{k}^*$, as the training exemplar.

Lexical Condition Lexically CPG uses pre-specified keywords to constrain paraphrasing, which requires a paraphrasing dataset containing ⟨*sentence*, *keywords*, *paraphrase*⟩ triples. Because the original dataset is formatted as ⟨*sentence*, *paraphrase*⟩, we need to pre-specify keywords for each data item. Following Zeng et al. (2019), we automatically extract keywords from the target sentence as the condition in the training stage. Besides, as also lacking manual keywords for each testing pair, we carry out two strategies for inference. On the one hand, we directly extract keywords from references as conditions following Zeng et al. (2019). On another, a standard sequence-to-sequence model is used to predict target keywords only from source sentences as conditions while testing, as described in Liu et al.(2020b). Specifically, we investigate three representative keyword extraction methods to verify the effectiveness of GCPG, including rule-based TF-IDF, TextRank (Mihalcea and Tarau, 2004), and model-based KeyBERT (Grootendorst, 2020). Each method filters out the stop words and punctuation, and guarantees the extracted keywords do not appear in the corresponding source sentence. The maximum number of keywords is set to 3. Besides, we use a special token "[NONE]" when there are no keywords extracted.

4 Experiments

Datasets Following previous works (Kumar et al., 2020; Bui et al., 2021), we evaluate GCPG on two datasets: (1) **ParaNMT-small** (Chen et al., 2019) is a subset of ParaNMT-50M dataset (Wieting and Gimpel, 2018), which is collected via back-translation referring to English sentences. It contains 500K training pairs formatted as ⟨sentence, paraphrase⟩, and 1.3K manually labeled data triples formatted as ⟨*sentence*, *exemplar*, *paraphrase*⟩ (0.8K for testing and 0.5K for validation). In each triple, *exemplar* is a sentence that has the same syntax as *paraphrase* but is semantically different from *sentence*. (2) **QQP-Pos** (Kumar et al., 2020) is selected from Quora Question Pairs (QQP) dataset. It contains about 140K training pairs and 3K/3K data triples for testing/validation. The format of dataset is the same as ParaNMT-small.

4.1 Syntactically Controllable Paraphrasing

We explore four syntactical conditions reconstructed by GCPG on the ParaNMT-small dataset, then compare SSE with baselines on two datasets.

Baselines We first choose two direct return-input baselines as dataset quality indicators: (1) *Source-as-Output* copies inputs as outputs. (2) *Exemplar-as-Output* regards exemplars as outputs. Next, we compare GCPG with mainstream competitive models as follows. (3) *SCPN* (Iyyer et al., 2018) has two encoders to encode source sentence and LCT separately, then constrain generation with soft attention mechanism.[4] (4) *CGEN* (Chen et al., 2019) encodes exemplars into latent vector to guide paraphrasing.[5] (5) *SGCP* (Kumar et al., 2020) uses a graph encoder to process the exemplar Constituent Trees as the condition.[6] (6) *ParafraGPT* (Bui et al., 2021) masks words with certain POS types in the target sentence as condition, then builds a paraphrasing generator based on a pre-trained GPT2 (Radford et al., 2019).

Syntactical Conditions We first examine conditions with manufactured syntax features, including (7) *POS Sequence*, (8) *LCT-Truncated* is the *LCT* sequence without POS-level information, (9) *LCT* is the full-fledged Linearised Constituent Tree sequence, and (10) *Masked Template*. Then, two implementations of SSE are evaluated: (11) *SSE-POS Sequence* uses *POS Sequence* to measure syntax similarity, and (12) *SSE-LCT-Truncated* uses *LCT-Truncated* as measurement.

Implementation and Hyper-parameters All GCPG models are instantiated by ProphetNet-large (Qi et al., 2020)[7]. We employ the original hyper-parameter setting of ProphetNet-large to train GCPG. During inference, the beam size and length penalty(Wu et al., 2016) are set to 4 and 1.2 following Bui et al. (2021). As for the

[4]`https://github.com/miyyer/scpn`

[5]`https://github.com/mingdachen/syntactic-template-generation`

[6]`https://github.com/malllabiisc/SGCP`

[7]We compare performances of different generation models for backbone selecting, details can be found in Appendix A.

Model	iBLEU ↑	B-R ↑	R-1 / R-2 / R-L ↑	MTR ↑	BS ↑	TED ↓
ParaNMT-small						
(1) Source-as-Output	-17.05	18.50	23.10 / 47.70 / 12.00	28.80	86.20	12.00
(2) Exemplar-as-Output	2.31	3.30	24.40 / 7.50 / 29.10	12.10	74.20	5.90
(3) SCPN (2018)	–	6.40	30.30 / 11.20 / 34.60	14.60	73.70	9.10
(4) CGEN (2019)	8.14	13.60	44.80 / 21.00 / 48.30	24.80	79.50	**6.70**
(5) SGCP (2020)	6.95	16.40	49.60 / 22.90 / 50.50	27.20	80.50	6.80
(6) ParafraGPT (2021)	8.61	14.54	49.67 / 22.42 / 51.29	27.83	90.78	8.22
(7) GCPG (POS Sequence)	11.96	19.97	56.20 / 32.36 / 58.99	32.68	92.57	8.45
(8) GCPG (LCT-Truncated)	**12.74**	22.54	59.98 / 36.81 / 62.61	37.04	93.39	8.34
(9) GCPG (LCT)	11.92	19.52	55.75 / 30.54 / 58.88	31.35	92.42	7.84
(10) GCPG (Masked Template)	9.52	16.85	53.60 / 27.96 / 56.31	31.84	92.21	8.84
(11) GCPG (SSE-POS Sequence)	10.07	23.82	60.93 / 37.36 / 61.98	36.15	91.55	8.94
(12) GCPG (SSE-LCT-Truncated)	12.32	**26.24**	**63.62 / 40.76 / 64.98**	**39.79**	**93.86**	8.27
QQP-Pos						
(13) Source-as-Output	-17.96	17.20	51.90 / 26.20 / 52.90	31.10	84.90	16.20
(14) Exemplar-as-Output	10.64	16.80	38.20 / 20.50 / 43.20	17.60	78.20	4.80
(15) SCPN (2018)	–	15.60	40.60 / 20.50 / 44.60	19.60	77.60	9.10
(16) CGEN (2019)	17.60	29.94	58.53 / 37.42 / 61.74	32.90	92.82	6.43
(17) SGCP (2020)	19.97	38.00	68.10 / 45.70 / 70.20	41.30	94.53	6.80
(18) ParafraGPT (2021)	21.19	35.86	66.71 / 43.70 / 68.94	40.26	94.54	6.11
(19) GCPG (SSE-LCT-Truncated)	**22.64**	**42.88**	**72.26 / 51.34 / 74.22**	**46.63**	**95.86**	**5.31**

Table 2: Results of different syntactical conditions and comparisons with baselines on ParaNMT-small and QQP-Pos datasets. B-R: BLEU-R. R-1: ROUGE-1. R-2: ROUGE-2. R-L: ROUGE-L. MTR: METEOR. BS: BERTScore. ↑ means higher score is better where ↓ is exactly the opposite. The highest numbers are in **bold**.

SSE dictionary size, the ParaNMT-small has 27530 truncated LCT (the average number of sentences for each LCT: 2.83) and QQP-Pos has 8561 (the average number of sentences for each LCT: 5.99).

Metrics Following previous works (Iyyer et al., 2018; Bui et al., 2021), we evaluate generating results on six metrics, including BLEU-4 (Papineni et al., 2002), ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L) (Lin, 2004), Meteor (MTR) (Denkowski and Lavie, 2014), and BERTScore (BS) (Zhang et al., 2020). Besides, *Source-as-Output* will also get a high BLEU score and BERTScore, we introduce iBLEU (Sun and Zhou, 2012) for more precise evaluation. As a variant of BLEU, iBLEU considers both fidelity to *reference* and diversification from *input*:

$$\begin{aligned} \text{iBLEU} &= \alpha\text{BLEU-R} - (1 - \alpha\text{BLEU-S}), \\ \text{BLEU-R} &= \text{BLEU-4}\,(\textit{output}, \textit{reference}), \\ \text{BLEU-S} &= \text{BLEU-4}\,(\textit{output}, \textit{input}), \end{aligned} \tag{3}$$

where the constant α is set to 0.7, as in the original paper. Finally, for syntactical condition evaluation, we follow Kumar et al. (2020) to calculate Tree-Edit Distance (TED)[8] between the Constituency Parse Trees of both *output* and *reference*.

[8]We use the evaluation tool implemented by SGCP.

Results As shown in Table 2, the main conclusions are: (1) SSE consistently and significantly outperforms conditions that constructed with manufactured syntax features (Rows 11-12 vs. Rows 7-10). (2) GCPG with SSE gets significant improvement over the previous SOTA (Row 12/19 vs. Row 6/18). (3) All syntactical conditions reconstructed in GCPG outperform baselines (Rows 7-12 vs. Rows 3-6), demonstrating the superiority of GCPG paradigm. However, the TED of GCPG is lower than CGEN on ParaNMT-small dataset. As the ParaNMT-small contains various noise data points, it is optimistic to assume that the corresponding constituency parse tree could be well aligned (Kumar et al., 2020), which may limit TED in evaluation. To address this issue, we also conduct a human evaluation (details in § 4.3). As shown in Table 4, GCPG with SSE (GCPG-S) performs better than CGEN on Syntax score.

4.2 Lexically Controllable Paraphrasing

As mentioned in § 3.2, we use three different keyword extraction methods to pre-specify keywords and comprehensively evaluate the GCPG: (1) *TF-IDF* (2) *TextRank* (Mihalcea and Tarau, 2004), and (3) *KeyBERT* (Grootendorst, 2020). Also, we compare GCPG with recent competitive method CTRL-

Condition	iBLEU ↑	B-R ↑	R-1 / R-2 / R-L ↑	MTR ↑	BS ↑	TED ↓
Keywords Extraction, GCPG instantiated by ProphetNet						
(1) GCPG (None)	4.67	18.46	55.29 / 31.17 / 55.18	32.42	92.32	11.78
(2) CTRLsum (2020)*	10.06	21.38	58.04 / 35.63 / 58.62	35.41	92.62	11.13
(3) GCPG (TF-IDF*)	10.07	23.04	61.92 / 38.68 / 61.71	36.97	92.86	10.79
(4) GCPG (TextRank*)	8.16	19.63	56.04 / 32.08 / 56.54	33.60	92.45	12.47
(5) GCPG (KeyBERT*)	11.03	24.12	60.92 / 38.00 / 61.14	35.41	92.79	10.26
(6) GCPG (KeyBERT (Upper Bound))	16.06	28.64	67.81 / 43.99 / 66.30	40.27	93.44	9.98
Keywords (KeyBERT*) + Syntactical Condition, GCPG instantiated by ProphetNet						
(7) GCPG (KeyBERT* + POS Sequence)	15.10	25.22	62.96 / 39.04 / 65.32	36.42	90.96	8.01
(8) GCPG (KeyBERT* + LCT-Truncated)	15.38	26.80	66.07 / 43.52 / 68.07	39.53	90.56	8.08
(9) GCPG (KeyBERT* + LCT)	14.47	23.52	61.92 / 36.33 / 64.38	34.73	92.74	8.00
(10) GCPG (KeyBERT* + Mask Template)	12.13	20.98	58.83 / 33.58 / 61.01	35.02	92.67	8.44
(11) GCPG (KeyBERT* + SSE-POS)	15.67	**31.02**	66.85 / 45.30 / 68.48	40.12	90.39	7.95
(12) GCPG (KeyBERT* + SSE-LCT-Truncated)	**15.73**	30.92	**68.40 / 46.73 / 69.93**	**41.98**	**94.34**	7.95

Table 3: Performance of different conditions and combinations under GCPG on ParaNMT-small. For fair comparison, we use CTRLsum instantiated by ProphetNet. Result with * means that we use a vanilla ProphetNet trained on the same dataset to predict keywords for GCPG while testing.

sum (He et al., 2020), which extracts entity words as keywords. We follow the settings in § 4.1.

Metrics For lexical condition, it should be noted that there is a lack of the explicit request of desirable keywords in the testing set. A generated paraphrase hinted by model predicted keywords might get a low score in BLEU, although humans consider it reasonable. This is because paraphrasing models might focus on keywords that are not consistent with the single reference. Therefore, we evaluate GCPG in three settings. First, following Liu et al.(2020b), we use a keywords prediction model to generate top-k groups of keywords, which are fed into GCPG to generate k paraphrases. Then the sentence that has the highest BLEU with the reference is selected as the final output (marked with *). k is set to 4 as well as beam size. Note that we use this setting to report the final results unless otherwise specified. Second, we further conduct human evaluations on the keyword condition based on *KeyBERT* (The details are in § 4.3). We denote it as "GCPG-L (k=1)". Here "k=1" means GCPG only produces one paraphrase for each input, constrained by the top-1 set of keywords produced by *KeyBERT*. Third, following Zeng et al. (2019), we directly extract keywords from references as the condition, marked with "(Upper Bound)".

Results As shown in the first five rows of Table 3, *KeyBERT* outperforms other two keyword extraction methods and CTRLsum on the iBLEU score. Also, GCPG with keyword condition significantly performs better than GCPG without it, which verifies the lexically controllable ability of GCPG.

4.3 Combinations

To facilitate the description, we define that "GCPG-L" denotes GCPG with the keyword condition extracted by *KeyBERT*, "GCPG-S" is GCPG with the *SSE-LCT-Truncated* condition, and "GCPG-LS" indicates the combination of conditions in "GCPG-L" and "GCPG-S". Meanwhile, GCPG is also instantiated by ProphetNet-large.

Metrics We follow the metrics in § 4.1, yet the automatic evaluations can not fully capture the fluency and the quality of the generation results on CPG. Therefore, we conduct human evaluation following Kumar et al.(2020). Specifically, we evaluate GCPG with different conditions on ParaNMT-small, then choose the best settings to compare GCPG with baselines on QQP-Pos. 100 test samples are randomly selected from each dataset. Then, 5 crowdsource evaluators are shown a source sentence and the corresponding reference, then asked to rate model results in three categories: whether the paraphrase remains **loyalty** to the source sentence, the **fluency** of paraphrase, and **syntax** similarity with gold reference.[9] Scores are ranged from 1 to 4, and the higher score is better.

Results As shown in Table 3, combinations of lexical and syntactical conditions get consistently further improvements compared with employing lexical condition individually (Rows 7-12 vs. Row 5). Then, we illustrate human evaluations in Table 4. GCPG with lexical condition (GCPG-L (k=1)) outperforms baselines in meaning and fluency, yet poor in syntax similarity. More impor-

[9]Details can be found in Appendix C.

Model	Loyalty	Fluency	Syntax	All
ParaNMT-small				
CGEN (2019)	1.47	2.13	1.81	5.41
ParafraGPT (2021)	1.86	2.42	2.05	6.33
GCPG-L (k=1)	2.94	3.63	2.29	8.86
GCPG-S	2.87	3.36	2.57	8.80
GCPG-LS (k=1)	3.09	3.51	2.46	**9.06**
QQP-Pos				
CGEN (2019)	1.72	2.52	2.22	6.46
ParafraGPT (2021)	2.43	2.91	2.61	7.95
GCPG-LS (k=1)	2.97	3.43	2.81	**9.21**

Table 4: Results of Human evaluation.

Model	BLEU-Exemplar ↓	
	ParaNMT-small	QQP-Pos
ParafraGPT (2021)	7.32	24.31
GCPG-S	2.63	23.17
Reference	3.30	16.80

Table 5: GCPG can significantly reduce BLEU-Exemplar score compared with previous SOTA.

tantly, the combination of lexical and syntactical conditions (GCPG-LS (k=1)) shows significantly improvements on all three scores.

4.4 Analyses and Discussions

We conduct discussions to shed light on other interesting properties of GCPG. Due to space constraints, we take discussions with GCPG instantiated by ProphetNet-large.

Exemplar-side Words Copying Problem We calculate BLEU-4 between model outputs and exemplars. As shown in Table 5, GCPG with SSE (i.e., GCPG-S) can significantly reduce BLEU-Exemplar comparing with ParafraGPT, gets 4.69 / 1.14 improvements on two datasets, demonstrating that SSE effectively alleviates this problem.

Lexical Condition Analyze further investigate GCPG with lexical condition from two aspects: 1) Generating novel expressions; 2) How frequently and correctly do the model incorporate the lexical condition in outputs? For the first one, following Dou et al.(2021), the number of novel n-grams is counted in the model output. Specifically, these n-grams appear in gold references but not in source sentences. After normalized by the total number of n-grams, we calculate the recall of novel n-grams. As shown in Figure 4, GCPG indeed generates novel expressions. Then, we explore the lexical condition accuracy of the settings in Table 3 (Row 2 vs Row 5), i.e., whether offered keywords appear in final outputs. Specifically, we ignore the model outputs that only use "[NONE]" as the keyword while paraphrasing for a fair comparison. As shown in Table 6, GCPG-L outperforms CTRLsum (He et al., 2020) with 9.8% improvement.

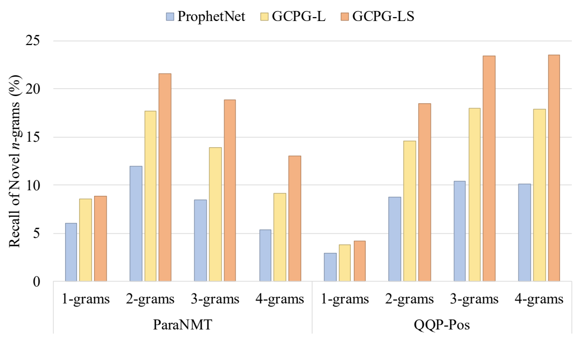


Figure 4: Recall of novel n-grams results.

Method	Accuracy (%)	iBLEU ↑
CTRLsum (2020)	89.45	10.06
GCPG-L	99.25	11.03

Table 6: Lexical condition accuracy on ParaNMT-small.

GCPG Backbone Analyze Whether the performance gain only from using strong PLMs is also of concern. As shown in Table 7, the results show that GCPG instantiated by vanilla Transformer gets comparable performance with ParafraGPT[10]. Also, we compare GCPG with ParafraGPT instantiated by the same backbone. GCPG still outperforms ParafraGPT with 2.8 iBLEU score.

Backbone	Method	iBLEU ↑
Transformer (2017)	-	4.72
BART (2020)	-	6.08
ProphetNet (2020)	-	4.67
GPT2 (2019)	ParafraGPT (2021)	8.61
ProphetNet (2020)	ParafraGPT‡	9.52
Transformer (2017)	GCPG-S	8.53
BART (2020)	GCPG-S	10.12
ProphetNet (2020)	GCPG-S	12.32

Table 7: Performance of GCPG-S with different backbones on ParaNMT-small. ‡: for fair comparisons, ParafraGPT is instantiated by ProphetNet.

5 Conclusions

In this paper, we propose a general framework GCPG, enabling flexibly combine lexical and syn-

[10]Due to space constraints, the results of all evaluations can be found in Appendix D.

tactical conditions and exploring their mutual effectiveness. Under GCPG, we provide SSE that allows GCPG to directly model syntax information from natural sentences and better utilize PLMs. As we tentatively give a successful implementation of leveraging two types of conditions in a unified circumstance, multilingual CPG and more types of conditions are barely being discussed. In the future, we will investigate to uniformly represent these conditions in a more superior way.

References

Tien-Cuong Bui, Van-Duc Le, Hai-Thien To, and Sang-Kyun Cha. 2021. Generative pre-training for paraphrase generation by representing and predicting spans in exemplars. In *IEEE BigComp*, pages 83–90. IEEE.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. In *ACL*, pages 5972–5984. ACL.

Douglass Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. 1992. A practical part-of-speech tagger. In *Third Conference on Applied Natural Language Processing*, pages 133–140.

Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *WMT-ACL*, pages 376–380. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186. ACL.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. Gsum: A general framework for guided neural abstractive summarization. In *NAACL*, pages 4830–4842. ACL.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *ACL*, pages 6065–6075. ACL.

Tanya Goyal and Greg Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. In *ACL*, pages 238–252. ACL.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Yunfan Gu, Yang Yuqiao, and Zhongyu Wei. 2019. Extract, transform and filling: A pipeline model for question paraphrasing based on template. In *W-NUT*, pages 109–114.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Fatema Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *CoRR*, abs/2012.04281.

Chris Hokamp and Qun Liu. 2017a. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546.

Chris Hokamp and Qun Liu. 2017b. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1535–1546. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *NAACL*, pages 1875–1885. ACL.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha P. Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Trans. Assoc. Comput. Linguistics*, 8:330–345.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880. ACL.

Yinghao Li, Rui Feng, Isaac Rehg, and Chao Zhang. 2020. Transformer-based neural text generation with syntactic guidance. *CoRR*, abs/2010.01737.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. ACL.

Dayiheng Liu, Jie Fu, Qian Qu, and Jiancheng Lv. 2019. Bfgan: backward and forward generative adversarial networks for lexically constrained sentence generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):2350–2361.

Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. 2020a. Revision in continuous space: Unsupervised text style transfer without adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8376–8383.

Dayiheng Liu, Yeyun Gong, Yu Yan, Jie Fu, Bo Shao, Daxin Jiang, Jiancheng Lv, and Nan Duan. 2020b. Diverse, controllable, and keyphrase-aware: A corpus and method for news multi-headline generation. In *EMNLP*, pages 6241–6250. ACL.

Dayiheng Liu, Yu Yan, Yeyun Gong, Weizhen Qi, Hang Zhang, Jian Jiao, Weizhu Chen, Jie Fu, Linjun Shou, Ming Gong, Pengcheng Wang, Jiusheng Chen, Daxin Jiang, Jiancheng Lv, Ruofei Zhang, Winnie Wu, Ming Zhou, and Nan Duan. 2021. GLGE: A new general language generation evaluation benchmark. In *Findings of ACL*, pages 408–420. ACL.

Mingtong Liu, Erguang Yang, Deyi Xiong, Yujie Zhang, Chen Sheng, Changjian Hu, Jinan Xu, and Yufeng Chen. 2020c. Exploring bilingual parallel corpora for syntactically controllable paraphrase generation. In *IJCAI*, pages 3955–3961. ijcai.org.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Comput. Linguistics*, 36(3):341–387.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *EACL*, pages 881–893. ACL.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2020. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Kathleen R. McKeown. 1983. Paraphrasing questions using given and new information. *Am. J. Comput. Linguistics*, 9(1):1–10.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *EMNLP*, pages 404–411.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1314–1324. Association for Computational Linguistics.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In *EMNLP*, volume EMNLP 2020 of *Findings of ACL*, pages 2401–2410. ACL.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Cynthia M Shewan. 1985. Auditory comprehension problems in adult aphasic individuals. *Human Communication Canada*, 9(5):151–155.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *ACL*, pages 38–42. ACL.

Jiao Sun, Xuezhe Ma, and Nanyun Peng. 2021. AESOP: paraphrase generation with adaptive syntactic control. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5176–5189. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

John Wieting and Kevin Gimpel. 2018. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *ACL*, pages 451–462. Association for Computational Linguistics.

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

Kexin Yang, Wenqiang Lei, Dayiheng Liu, Weizhen Qi, and Jiancheng Lv. 2021. Pos-constrained parallel decoding for non-autoregressive generation. In *ACL*, pages 5990–6000. ACL.

Daojian Zeng, Haoran Zhang, Lingyun Xiang, Jin Wang, and Guoliang Ji. 2019. User-oriented paraphrase generation with keywords controlled network. *IEEE Access*, 7:80542–80551.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*. OpenReview.net.

A GCPG Backbone Selection

We evaluate the following text generation models to select a backbone model for our GCPG. (1) *Transformer* (Vaswani et al., 2017), the conventional version in the original paper. (2) *BART* (Lewis et al., 2020)[11] has a denoising autoencoder for pre-training sequence-to-sequence models, and (3) ProphetNet-large (Qi et al., 2020)[12], which has shown its effectiveness in text generation (Liu et al., 2021). The results are shown in Table 8.

[11] `https://github.com/pytorch/fairseq/tree/master/examples/bart`

[12] `https://github.com/microsoft/ProphetNet`

Model	iBLEU ↑	B-R ↑	R-1 / R-2 / R-L ↑	MTR ↑	BS ↑	TED ↓
ParaNMT-small						
Transformer	4.72	14.66	51.05 / 26.88 / 51.32	30.67	91.30	12.71
BART	6.08	17.78	52.37 / 27.02 / 51.52	31.57	91.99	11.92
ProphetNet	4.67	18.46	55.29 / 31.17 / 55.18	32.42	92.32	11.78
QQP-Pos						
Transformer	7.63	23.44	54.58 / 30.48 / 56.63	32.60	93.18	11.84
BART	3.14	23.07	56.43 / 32.12 / 57.64	34.26	93.58	13.05
ProphetNet	6.43	25.79	58.40 / 34.52 / 59.98	35.75	93.88	11.74

Table 8: Results of different generation models on ParaNMT-small and QQP-Pos datasets. B-R: BLEU-R. R-1: ROUGE-1. R-2: ROUGE-2. R-L: ROUGE-L. MTR: METEOR. BS: BERTScore. ↑ means higher score is better where ↓ is exactly the opposite. The highest numbers are in **bold**.

B GCPG with syntactical conditions on QQP-Pos dataset

The experimental results can be found in Table 9.

C Human Evaluation Details

For human evaluation, we first set a guideline for evaluating, which includes the task background, key points, detailed descriptions and examples of evaluation scores from 1 to 4. Then, we set an entry barrier for annotators. In detail, we organize a training program and a preliminary annotating examination (50 examples for each data set) to select appropriate annotators.

Score Definition We define three categories in human evaluation as follows:

Loyalty means whether the option is consistent with the meaning of the original sentence (whether the content is missing or omitted).

Fluency means whether the sentence corresponding to the option is fluent.

Syntax is the similarity of grammatical structure with reference answer, which means whether the sentence structure of this option similar to the reference answer.

Inter-annotator agreement We use Fleiss' kappa (Fleiss, 1971) to measure 5 annotator's reliability[13]. The results are : 1) Paranmt-small dataset (loyalty: 0.56, fluency: 0.42, syntax: 0.41); 2) QQP-Pos (loyalty: 0.55, fluency: 0.40, syntax: 0.37).

D GCPG Backbone Analyze

All kinds of scores in GCPG Backbone Analyze can be found in Table 10.

E Case Study

The qualitative effect of the lexical and syntactical conditions on the model output is also of interest. To intuitively display the effects of conditions, we show some paraphrasing results in Figure 5. In detail, GCPG-L can generate sentence "A powerful healing energy comes out of love." that contain pre-specified keywords "[healing]". However, lexical condition provides less information about syntactical controlling. In comparison, GCPG-LS shows better performances on both controllability of lexical items and syntax.

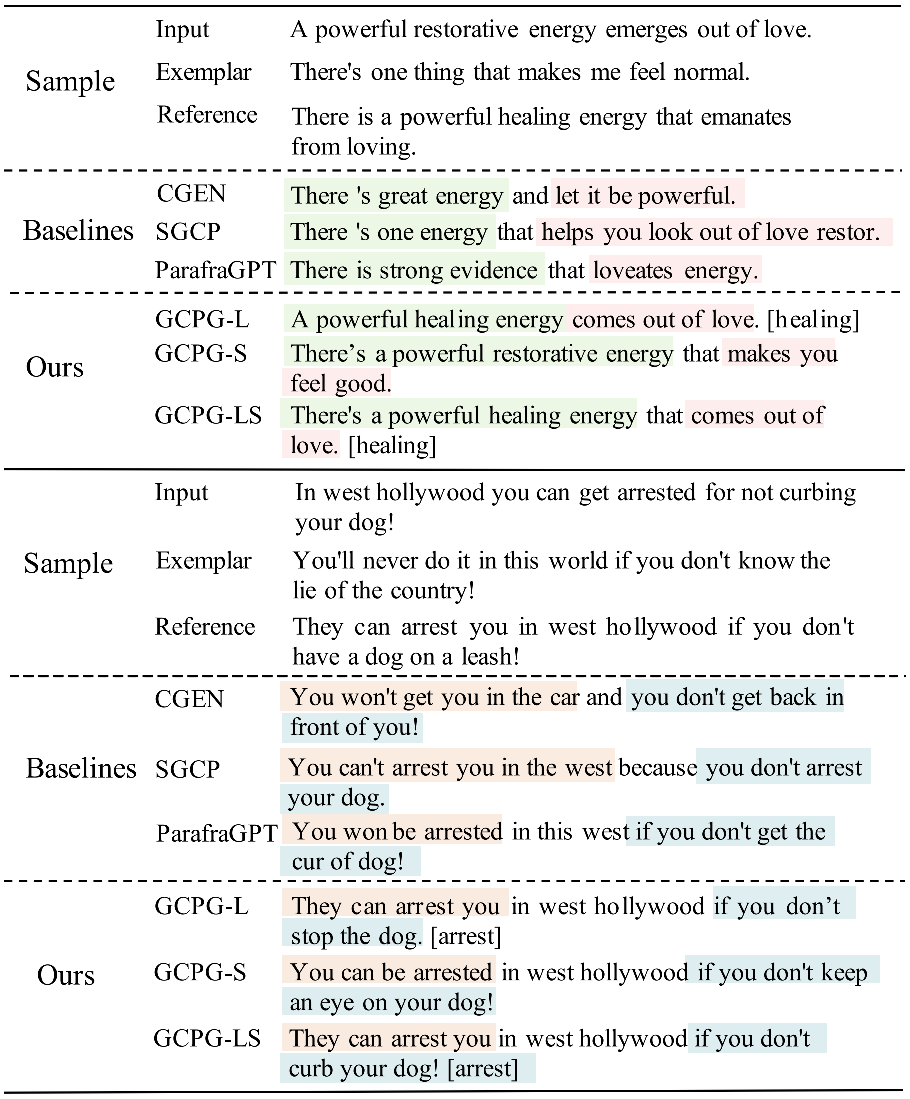
[13]https://www.nltk.org/_modules/nltk/metrics/agreement.html

Model	**iBLEU ↑**	**B-R ↑**	**R-1 / R-2 / R-L ↑**	**MTR ↑**	**BS ↑**	**TED ↓**
			QQP-Pos			
(1) GCPG (POS Sequence)	17.84	37.49	70.10 / 47.43 / 71.49	43.05	95.46	7.04
(2) GCPG (LCT-Truncated)	20.93	40.55	71.31 / 49.20 / 73.31	45.17	95.83	5.66
(3) GCPG (LCT)	20.17	38.88	70.61 / 48.00 / 72.49	42.95	95.58	6.11
(4) GCPG (Masked Template)	19.16	33.65	64.78 / 42.17 / 67.31	38.97	94.86	6.22
(5) GCPG (SSE-POS Sequence)	21.56	42.63	**72.78 / 52.92 / 74.53**	**47.08**	**95.88**	5.85
(6) GCPG (SSE-LCT-Truncated)	**22.64**	**42.88**	72.26 / 51.34 / 74.22	46.63	95.86	**5.31**

Table 9: Results of different syntactical conditions on QQP-Pos dataset. B-R: BLEU-R. R-1: ROUGE-1. R-2: ROUGE-2. R-L: ROUGE-L. MTR: METEOR. BS:BERTScore. ↑ means higher score is better where ↓ is exactly the opposite. The highest numbers are in **bold**.

Backbone	**Method**	**iBLEU ↑**	**B-R ↑**	**R-1 / R-2 / R-L ↑**	**MTR ↑**	**BS ↑**	**TED ↓**
Transformer (2017)	-	4.72	14.66	51.05 / 26.88 / 51.32	30.67	91.30	12.71
BART (2020)	-	6.08	17.78	52.37 / 27.02 / 51.52	31.57	91.99	11.92
ProphetNet (2020)	-	4.67	18.46	55.29 / 31.17 / 55.18	32.42	92.32	11.78
GPT2 (2019)	ParafraGPT (2021)	8.61	14.54	49.67 / 22.42 / 51.29	27.83	90.78	8.22
ProphetNet (2020)	ParafraGPT‡	9.52	16.85	53.60 / 27.96 / 56.31	31.84	92.21	8.84
Transformer (2017)	GCPG-S	8.53	17.14	55.89 / 30.83 / 57.15	33.29	92.24	9.67
BART (2020)	GCPG-S	10.12	19.08	57.87 / 34.65 / 59.83	34.48	92.50	11.02
ProphetNet (2020)	GCPG-S	12.32	26.24	63.62 / 40.76 / 64.98	39.79	93.86	8.27

Table 10: Performance of different Backbones under GCPG-S on ParaNMT-small. ‡: for fair comparisons, ParafraGPT (previous SOTA) is also instantiated by ProphetNet.



Sample	Input	A powerful restorative energy emerges out of love.
	Exemplar	There's one thing that makes me feel normal.
	Reference	There is a powerful healing energy that emanates from loving.
Baselines	CGEN	There 's great energy and let it be powerful.
	SGCP	There 's one energy that helps you look out of love restor.
	ParafraGPT	There is strong evidence that loveates energy.
Ours	GCPG-L	A powerful healing energy comes out of love. [healing]
	GCPG-S	There’s a powerful restorative energy that makes you feel good.
	GCPG-LS	There's a powerful healing energy that comes out of love. [healing]
Sample	Input	In west hollywood you can get arrested for not curbing your dog!
	Exemplar	You'll never do it in this world if you don't know the lie of the country!
	Reference	They can arrest you in west hollywood if you don't have a dog on a leash!
Baselines	CGEN	You won't get you in the car and you don't get back in front of you!
	SGCP	You can't arrest you in the west because you don't arrest your dog.
	ParafraGPT	You won be arrested in this west if you don't get the cur of dog!
Ours	GCPG-L	They can arrest you in west hollywood if you don’t stop the dog. [arrest]
	GCPG-S	You can be arrested in west hollywood if you don't keep an eye on your dog!
	GCPG-LS	They can arrest you in west hollywood if you don't curb your dog! [arrest]

Figure 5: Samples of paraphrases. Words in “[]” are offered by our keywords prediction model based on KeyBERT. We highlight different parts for better view.