Toward the Modular Training of Controlled Paraphrase Adapters

Teemu Vahtola and **Mathias Creutz**
Department of Digital Humanities
Faculty of Arts
University of Helsinki
Finland
{teemu.vahtola, mathias.creutz}@helsinki.fi

Abstract

Controlled paraphrase generation often focuses on a specific aspect of paraphrasing, for instance syntactically controlled paraphrase generation. However, these models face a limitation: they lack modularity. Consequently adapting them for another aspect, such as lexical variation, needs full retraining of the model each time. To enhance the flexibility in training controlled paraphrase models, our proposition involves incrementally training a modularized system for controlled paraphrase generation for English. We start by fine-tuning a pretrained language model to learn the broad task of paraphrase generation, generally emphasizing meaning preservation and surface form variation. Subsequently, we train a specialized sub-task adapter with limited sub-task specific training data. We can then leverage this adapter in guiding the paraphrase generation process toward a desired output aligning with the distinctive features within the sub-task training data.

The preliminary results on comparing the fine-tuned and adapted model against various competing systems indicates that the most successful method for mastering both general paraphrasing skills and task-specific expertise follows a two-stage approach. This approach involves starting with the initial fine-tuning of a generic paraphrase model and subsequently tailoring it for the specific sub-task.

1 Introduction

Paraphrase generation aims to produce sentences that maintain high semantic similarity with the source sentence, while deviating enough from it on surface form. Commonly used sequence-to-sequence models encounter challenges in generating diverse paraphrase outputs (Kumar et al., 2019). As a result, recent research in paraphrase generation has shifted toward controlled generation methods. These approaches condition the model on predefined qualities to produce specific outputs, aiming to overcome this limitation. Exploring approaches to controlled text generation has both theoretical and practical implications. It can influence the theoretical understanding of automatic language generation and offer practical applications across various domains and industries.

Through leveraging controlled text generation, models can for instance be steered to produce language that better follows user preferences (Fan et al., 2018). With enough surface form variation, paraphrasing can be useful in question answering (Dong et al., 2017), data augmentation (Kumar et al., 2019), and machine translation (Callison-Burch et al., 2006; Mehdizadeh Seraj et al., 2015), among other tasks. Even if trained to perform certain paraphrase transformations, recent controlled paraphrase generation systems are limited in flexibility. Incorporating an additional control feature necessitates retraining the entire model. To overcome this limitation, we make the assumption that paraphrase generation essentially behaves in a modular manner. To evaluate our assumption, we propose the training of a modular system for controlled paraphrase generation through initial fine-tuning or broader task adapters (Pfeiffer et al., 2020b) followed by more specialized sub-task adapters. Hence, in contrast to standard fine-tuning of all parameters of a model, we initially train the model to perform the necessary paraphrasing skills, namely meaning preservation and surface form variation, and further refine the model in a modular way to produce outputs that encompass some desired paraphrase nuances. We focus on English paraphrasing, incorporating one specific aspect of paraphrasing, namely antonym substitution (Bhagat and Hovy, 2013). We select this paraphrase operation due to the availability of a specialized test suite designed for evaluating paraphrase models on sentence pairs that incorporate antonym substitution (Vahtola et al., 2022), enabling systematic comparison of various experimental setups. However,

Proceedings of the 1st Workshop on Modular and Open Multilingual NLP (MOOMIN 2024), pages 1–6
March 21, 2024 ©2024 Association for Computational Linguistics

our proposed approach could as well be applied to other paraphrase phenomena and languages where paraphrase data is available.

2 Previous Research

Common methods for automatic paraphrase generation rely on sequence-to-sequence modeling, often leveraging machine translation (Tiedemann and Scherrer, 2019; Thompson and Post, 2020; Sun et al., 2022, *inter alia*), or monolingual parallel data (Prakash et al., 2016; Sjöblom et al., 2020). These models, however, often struggle with generating sufficient variation (Kumar et al., 2020). As a result, increased emphasis has been given to generating controlled paraphrases, specifically targeting variations across predefined dimensions.

There has been significant research attention directed toward controlled paraphrasing in various granularity levels, from aiming to produce lexical variation by providing synonym substitutions (Fu et al., 2019) to the generation of syntactically controlled paraphrases (Iyyer et al., 2018; Kumar et al., 2020; Sun et al., 2021). While these approaches are constrained by concentrating solely on one level of detail, diverse paraphrasing encompasses multiple levels of granularity. To acknowledge this limitation, Huang et al. (2019) use dictionaries to perform word-level and phrase-level paraphrasing, obtaining more variation. Vahtola et al. (2023) train a multilingual NMT model with control tokens related to various aspects of paraphrasing. It is still however an open question how the control tokens interplay. In addition, a critical limitation arises with these models: they lack modularity, wherein all control tokens exert simultaneous influence on the output, making it impossible to selectively deactivate any subset of control features during the inference process or flexibly adapt the model to new features. In contrast, we propose the training of a modular controlled paraphrase generation model leveraging adapter transformations (Houlsby et al., 2019; Pfeiffer et al., 2020b).

In addition to being widely studied for cross-lingual transfer (e.g., Pfeiffer et al., 2020b) and NMT (Üstün et al., 2021), modular and parameter efficient fine-tuning has been explored in other sequence-to-sequence tasks. Bapna and Firat (2019) use a modification of trainable adapter blocks (Houlsby et al., 2019) to adapt MT outputs for new languages and domains. Wan et al. (2023) leverage prefix-tuning for generating syntactically controlled paraphrases. In contrast to the previous work on modular fine-tuning, our focus lies in the modular training paradigm specific to paraphrasing. We delve into training specialized sub-task adapters within this single task. These adapters are supposed to capture task and sub-task specific information, and are to be assembled to produce controlled paraphrasing toward an intended output.

3 Data

We use the English partition of the Opusparcus paraphrase dataset (Creutz, 2018) for alternately fine-tuning the full model or training a generic paraphrase adapter. The training data in Opusparcus was automatically constructed and organized to prioritize the most probable paraphrastic sentence pairs at the beginning, with decreasing likelihood of being paraphrases as the data progresses. Hence, we select the first 1 000 000 sentence pairs from the corpus as training data, denoted as T, comprising approximately of 95% of true paraphrases (Creutz, 2018), and use the sentence pairs annotated as paraphrases from the Opusparcus development set for tuning the models. Moreover, within the training set T, we extract a specialized subset $T_n \subset T$ consisting of 12 870 examples. We use the first 12 000 examples as training data and save the final 870 examples to serve as a development set for tuning the specialized systems. This subset exclusively comprises instances where an explicit negation token is present in the target but absent in the source, and is used for training a dedicated sub-task adapter as a part of a broader paraphrasing task. We aim to extract sentences that demonstrate interesting paraphrastic relationships through the use of negation or negated antonymy, as opposed to sentences that negate the intended meaning. We release the task-specific data in `https://github.com/teemuvh/controlled-paraphrase-adapters`.

4 Experiments

Our objective is to incrementally train and assemble a modular system for controlled paraphrase generation. We undertake training and assessment across several models. To start, we establish a baseline by fine-tuning `flan-t5-base`[1] (Chung et al., 2022) using a set of 1 000 000 paraphrase pairs (T) sourced from the English partition of the Opusparcus training set. Furthermore, we fine-tune a sepa-

[1] The prefix we use for training and evaluating the models is: *paraphrase this sentence:*.

Original	Candidates
You’re not fat.	You’re not thin., You’re fat., **You’re thin.**
It’s not fair.	It’s not unfair., It’s fair., **It’s unfair.**
This is not a good idea.	This is not a bad idea., This is a good idea., **This is a bad idea.**
It is not safe.	It is not dangerous., It is safe., **It is dangerous.**

Table 1: Examples from the SemAntoNeg test suite. The true paraphrase to the input sentence is highlighted.

rate system using only a subset of the training data (T_n) that comprises of examples incorporating paraphrasing through negation and negated antonymy, extracted from the complete training set. We also perform a two-stage fine-tuning, starting with fine-tuning the base model with T, and sequentially fine-tuning with T_n.

In all adapter experiments, we leverage the adapter-transformers library (Pfeiffer et al., 2020a). We optimize modular fine-tuning by utilizing the bottleneck adapter (Houlsby et al., 2019) configuration proposed in Pfeiffer et al. (2020b) in conjunction with the base model. We then proceed to train two task adapters: one using the entire training dataset (T) for a broad paraphrasing task, and another using a subset (T_n) of the data for a specific controlled paraphrasing sub-task. Finally, we explore incremental adapter training by enhancing the base model with the paraphrase adapter. We then freeze the weights of the base model and the paraphrase adapter and proceed to train an additional sub-task adapter. This adapter not only benefits from the paraphrase adapter’s information but also focuses on learning more specific paraphrasing transformations incorporating negation and antonym substitution. We train each system on a single GPU for 3 epochs with a batch size of 128, and 5e-5 learning rate.

We evaluate the models on a dedicated test suite designed for paraphrase detection within sequences incorporating negated antonyms (Vahtola et al., 2022). The test suite is intended to be used to evaluate models on a difficult paraphrase detection task involving sequences with high lexical overlap. Examples of the data are provided in Table 1. To make the test suite suitable for evaluating sequence-to-sequence models, we extract each source sentence and its true paraphrase, i.e., the third candidate as highlighted in the examples in Table 1, from the test suite. By treating these extracted pairs as source-target sequences, we reframe the task as a sequence-to-sequence challenge. A successful model hence performs antonym substitution to produce a paraphrase of the original sentence. Controlled paraphrasing aims to replicate a specific output sentence while incorporating predefined control features. Therefore, we decide to evaluate the models using BLEU (Papineni et al., 2002) with respect to the references and to the inputs. We use sacreBLEU (Post, 2018) for calculating the BLEU scores.

5 Results

Table 2 presents the results. The base model evaluation (denoted as base in Table 2), conducted without any fine-tuning or adaptation, establishes a baseline BLEU score of 25.07. Fine-tuning (para-ft) or training an adapter (para-adapt) solely with the 1 000 000 examples (T from now on) yields suboptimal results (14–17 BLEU) on the negated antonym test data. However, this outcome is expected, as the model is not explicitly trained to handle paraphrases with negation or negated antonyms. While the BLEU score may be lower for the paraphrase models, it doesn’t necessarily imply inferiority in their ability to paraphrase. As indicated by the high BLEU score with respect to the source sentence (S-BLEU in Table 2), the base model without fine-tuning or adapter training has a high tendency to copy the input sentence, consequently yielding relatively high BLEU score in this task owing to the extensive lexical overlap found within the test data examples. The dedicated paraphrase models aim to introduce more alternations to the inputs, resulting in lower BLEU scores despite potentially producing true paraphrases.

Fully fine-tuning the model with the filtered subset (T_n) of the training data (neg-ft), thus highlighting paraphrasing through negation and antonymy, consistently produces higher BLEU scores on the task compared to both the base model and models trained solely on T. Adapter training on top of the base model using T_n (neg-adapt) results in even higher BLEU scores. Parameter efficient fine-tuning has been shown to be effective in low-resource scenarios (e.g., Karimi Mahabadi

Model	BLEU	S-BLEU
base	25.07	95.23
para-ft	14.21	30.73
para-adapt	17.15	47.71
neg-ft	30.24	49.15
neg-adapt	32.79	57.92
para-ft+neg-ft	23.40	24.83
para-adapt+neg-adapt	26.06	36.45
para-ft+neg-adapt	34.00	66.45

Table 2: Results of the different models on the SemAntoNeg challenge set framed as a sequence-to-sequence task. Here, BLEU scores measure the alignment with reference sentences, whereas S-BLEU assesses alignment with the input itself.

et al., 2021), which might explain why the adapter method achieves higher BLEU scores compared to full fine-tuning when trained specifically for the given paraphrasing sub-task.

Initiating training by fine-tuning a generic paraphrase model, followed by further fine-tuning with the specific sub-task data yields a subpar model (para-ft+neg-ft). Similarly, training an extensive paraphrase adapter before introducing a specialized sub-task adapter (para-adapt+neg-adapt) results in a model which barely surpasses the base model's performance when evaluated against the reference using BLEU. Comparing the outputs to the input sentences however shows that the incrementally adapted model achieves similar BLEU scores as the base model by trying to produce variation rather than simply duplicating the input sentence, as indicated by the lower S-BLEU score of the adapted model.

The best BLEU scores are obtained by fully fine-tuning the base model leveraging all 1 000 000 paraphrase examples and training a specialized sub-task adapter on top of the refined model (para-ft+neg-adapt). We hypothesize that the initial fine-tuning steers the model toward generating outputs that highly resemble the input, reflected in a relatively high S-BLEU. Subsequent adapter training on a smaller scale then refines the model's proficiency in paraphrase operations involving negation and negated antonyms, as indicated by the highest BLEU.

The relationship between the obtained BLEU and S-BLEU is presented in Figure 1. A robust paraphrase model would typically demonstrate a balance between a higher BLEU score and a lower S-BLEU score, positioning itself toward the lower right corner of the diagram. This would indicate robustness by demonstrating a substantial lexical similarity between the input and the reference, while having a lesser alignment with the input itself. In our task, an oracle model producing the exact reference sentence would obtain 100 BLEU and 24.90 S-BLEU.

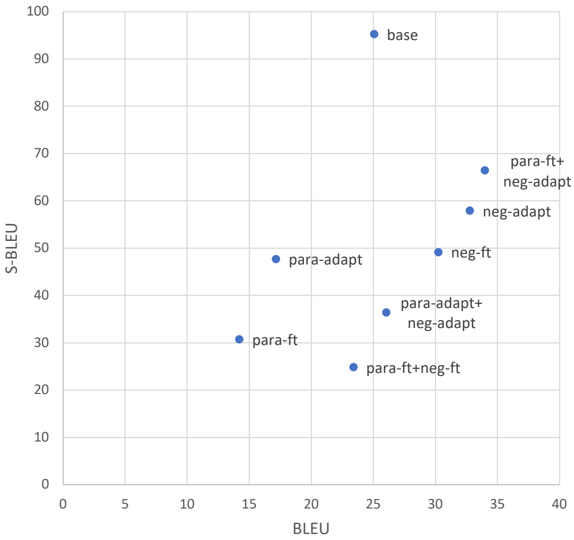

Figure 1: The BLEU and S-BLEU values of the methods shown graphically. The best performing models are assumed to show far to the right, reflecting a high BLEU with respect to the reference, and at around 25 % S-BLEU, which is the BLEU value of the reference with respect to the source. That is, an oracle model that would produce the desired reference sentences would obtain BLEU = 100 % and S-BLEU = 24.90 %.

To summarize the results, the base model along with the models subjected to plain fine-tuning or adaptation with the more generic paraphrase data exhibit poor performance, highlighted by the base model's high S-BLEU, and the low BLEU scores achieved by the fine-tuned or adapted models. Incorporating specialized training for the intended paraphrasing task, either through fine-tuning or adaptation, is essential for success in the task. However, the results obtained with the models specifically trained for paraphrasing through negation or negated antonymy remain somewhat inconclusive. Further analysis is necessary to determine the optimal training configuration for assembling general paraphrasing capabilities with specialized sub-task capabilities. Additionally, we hypothesize that parameter-efficient fine-tuning is better suited in scenarios involving limited data. However, the limited training data is also more task-specific, so it is still too early to draw general conclusions.

6 Conclusions

We propose the training of a modular paraphrase generation model that is built incrementally. This model starts by fine-tuning on a robust pretrained language model to learn the general requirements of paraphrase generation, namely meaning preservation and surface form variation. Subsequently, we train a specialized sub-task adapter with a limited number of sub-task specific training data to guide the paraphrase generation process toward a desired output. We compare the model involving fine-tuning followed by sub-task adaptation to several counterparts, including a base model without further training, as well as differently fine-tuned or adapted systems.

When assessing on a dedicated test set involving paraphrasing with negation or negated antonyms, we find that the most effective approach for learning both general paraphrasing abilities and sub-task specific expertise is achieved by fully fine-tuning a model for paraphrasing and then tailoring it to the specific sub-task through modular updates.

In future work, we wish to delve deeper into modularity for controlled paraphrasing. We intend to expand the model's capabilities by incrementally training it to encompass additional paraphrasing nuances, such as syntactic or lexical variation. Furthermore, we would like to assess how varying the size and task-specificity of the training data impacts the results. Finally, we would like to extend our approach to a multilingual setup.

Acknowledgements

This study was supported by the Behind the Words project, funded by the Research Council of Finland. We wish to acknowledge CSC – The Finnish IT Center for Science for the computational resources they have generously provided. We express our gratitude to Ivan Vulić for engaging in insightful discussions and providing valuable feedback during the conceptualization of this work.

References

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for 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 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 17–24, New York City, USA. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886, Copenhagen, Denmark. Association for Computational Linguistics.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Yao Fu, Yansong Feng, and John P Cunningham. 2019. Paraphrase generation with latent bag of words. *Advances in Neural Information Processing Systems*, 32.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Shaohan Huang, Yu Wu, Furu Wei, and Zhongzhi Luan. 2019. Dictionary-guided editing networks for paraphrase generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6546–6553.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. *Advances in Neural Information Processing Systems*, 34:1022–1035.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:329–345.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. 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 3609–3619, Minneapolis, Minnesota. Association for Computational Linguistics.

Ramtin Mehdizadeh Seraj, Maryam Siahbani, and Anoop Sarkar. 2015. Improving statistical machine translation with a multilingual paraphrase database. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1379–1390, Lisbon, Portugal. Association for Computational Linguistics.

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

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

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

Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual LSTM networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2923–2934, Osaka, Japan. The COLING 2016 Organizing Committee.

Eetu Sjöblom, Mathias Creutz, and Yves Scherrer. 2020. Paraphrase generation and evaluation on colloquial-style sentences. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1814–1822, Marseille, France. European Language Resources Association.

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*, pages 5176–5189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaofei Sun, Yufei Tian, Yuxian Meng, Nanyun Peng, Fei Wu, Jiwei Li, and Chun Fan. 2022. Paraphrase generation as unsupervised machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6379–6391, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Brian Thompson and Matt Post. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2019. Measuring semantic abstraction of multilingual NMT with paraphrase recognition and generation tasks. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 35–42, Minneapolis, USA. Association for Computational Linguistics.

Ahmet Üstün, Alexandre Berard, Laurent Besacier, and Matthias Gallé. 2021. Multilingual unsupervised neural machine translation with denoising adapters. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6650–6662, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Teemu Vahtola, Mathias Creutz, and Jörg Tiedemann. 2022. It is not easy to detect paraphrases: Analysing semantic similarity with antonyms and negation using the new SemAntoNeg benchmark. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 249–262, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Teemu Vahtola, Mathias Creutz, and Jrg Tiedemann. 2023. Guiding zero-shot paraphrase generation with fine-grained control tokens. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 323–337, Toronto, Canada. Association for Computational Linguistics.

Yixin Wan, Kuan-Hao Huang, and Kai-Wei Chang. 2023. PIP: Parse-instructed prefix for syntactically controlled paraphrase generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10372–10380, Toronto, Canada. Association for Computational Linguistics.