Supplementary Materials: A Contrastive Framework for Neural Text Generation

Appendix

Table of Contents

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

A Future Work

For future work, we would like to suggest three research directions based on our study.

- Our proposed contrastive loss $\mathcal{L}_{\text{CL}}$ in Eq. (2) is designed to treat all other tokens within the same sequence as negative samples. However, we do acknowledge that there might be a suitably small fraction of tokens (within the same sequence) that share similar semantic meanings even with different surface forms. We believe the current formulation of the contrastive loss might be further improved by taking this aspect into consideration and we leave it to our future work.
- One limitation of the proposed contrastive search is that it is a deterministic decoding method. It would be interesting and useful to incorporate a certain level of stochasticity into the decoding process. One plausible approach is to combine contrastive search with stochastic sampling methods. For instance, given the prefix, we could first generate a few tokens (e.g., 1$\sim$3 tokens) with nucleus sampling. Then, we switch to contrastive search for the remaining steps. In Appendix L, we provide some preliminary experiment results on incorporating stochasticity into contrastive search.
- Our approach is architecture agnostic and can be applied to any generation model. Future research could focus on adapting it to other tasks than open-ended text generation (i.e., constrained text generation), such as machine translation and document summarization.

B Related Work

Neural Text Generation is a core component in many NLP applications. It can be generally categorized into two classes (1) constrained generation; and (2) open-ended generation.

Constrained generation tasks are always defined over a set of (input, output) pairs, where the output is a transformation of the input following specific constrains. Some typical examples include machine translation [36, 2, 18], text summarization [24, 28], and data-to-text generation [39, 34, 32, 40]. As the output is tightly scoped by the input, the generation of repetition and unnaturalness are not that problematic, therefore maximization-based decoding methods such as beam search generally perform well. Still, different variants of beam search have been explored to further improve the model performance in constrained generation tasks [10, 11, 17, 16].

Open-ended text generation, on the other hand, imposes less constrain on the generated text. It aims at producing text that is natural, coherent and informative with respect to the human-written prefix (i.e., context). Several typical applications include story generation [6, 30], contextual text completion [23], and dialogue systems [35, 33]. However, due to the challenges posed by the increased level of freedom, conventional maximization-based decoding methods (e.g., greedy and beam search) often produce undesirable repetition and unnaturalness in the generated text. To alleviate model degeneration, different sampling approaches [6, 8, 19] have been proposed to generate text by drawing samples from less likely vocabularies. Welleck *et al.* [38] tackled model degeneration from another perspective by introducing unlikelihood objective into the training of the language model.

Contrastive Learning. Generally, contrastive learning methods aim to teach the model to distinguish observed data points from fictitious negative samples. They have been widely applied to various research areas. In the field of computer vision, contrastive learning has been shown to benefit tasks like image [37] and video [26] representation learning. Chen *et al.* [4] proposed a simple framework, SimCLR, for learning contrastive visual representations. Recently, Radford *et al.* [22] and Jia *et al.* [9] applied contrastive learning for the pre-training of language-image models.

In the field of NLP, contrastive learning has recently gained much more attention. Numerous contrastive approaches have been proposed to learn better token-level [31], sentence-level [20, 14, 7], and discourse-level [29, 13, 1, 12] representations. Beyond representation learning, contrastive learning has also been applied to other NLP applications, such as name entity recognition (NER) [5], document summarization [15], and knowledge probing for pre-trained language models [21].

Our work, to the best of our knowledge, is the first effort on applying contrastive learning to address neural text degeneration. We hope our findings could facilitate future research in this area.

C Software Package

In this section, we illustrate the use of the accompanying Python package, available on Github[1] and installable via pip[2] as `pip install simctg --upgrade`.

Below, we show how to replicate our result in Table 4 with our provided package. More details can be found in our open-sourced repository[3].

```
import torch
# load the language model
from simctg.simctggpt import SimCTGGPT
model_name = r'cambridgeltl/simctg_wikitext103'
model = SimCTGGPT(model_name)
model.eval()
tokenizer = model.tokenizer

# prepare input
prefix_text = # The prefix text in Table 4
print ('Prefix is: {}'.format(prefix_text))
tokens = tokenizer.tokenize(prefix_text)
input_ids = tokenizer.convert_tokens_to_ids(tokens)
input_ids = torch.LongTensor(input_ids).view(1,-1)

# generate result with contrastive search
beam_width, alpha, decoding_len = 8, 0.6, 128
output = model.fast_contrastive_search(input_ids=input_ids,
                         beam_width=beam_width, alpha=alpha,
                         decoding_len=decoding_len)
print("Output:\n" + 100 * '-')
print(tokenizer.decode(output))
```

Listing 1: Example usage of the SimCTG package

Model	Size	Objective	ppl↓	acc↑	conicity↓	self-similarity↓	Method	diversity↑	MAUVE↑	coherence↑
Transformers	117M	MLE	26.60	35.62	0.50	0.22	nucleus	0.89	0.81	0.541
							contrastive	0.90	0.83	0.561
		SimCTG	**26.55**	**36.03**	**0.47**	**0.19**	nucleus	0.89	0.82	0.543
							contrastive	**0.91**	**0.85**	**0.566**
GPT-2-small	117M	MLE	24.32	39.63	0.90	0.86	nucleus	0.94	0.90	0.577
							contrastive	0.24	0.18	0.599
		SimCTG	**23.82**	**40.91**	**0.43**	**0.18**	nucleus	0.94	0.92	0.584
							contrastive	**0.95**	**0.94**	**0.610**
GPT-2-large	774M	MLE	16.57	43.34	0.46	0.20	nucleus	0.94	0.91	0.583
							contrastive	**0.95**	**0.96**	0.623
		SimCTG	**16.53**	**43.47**	**0.42**	**0.17**	nucleus	**0.95**	0.93	0.591
							contrastive	**0.95**	**0.96**	**0.626**
Human	-	-	-	-	-	-	-	0.95	1.00	0.644

Table 1: Experimental results of different language models on Wikitext-103. ↑ means higher is better and ↓ means lower is better. The results of GPT-2-small are copied from Table 1.

D Experiments on Different Language Models

In this section, we further test the generalization ability of our approach with different language models on the Wikitext-103 benchmark. In addition to the GPT-2-small model (i.e. 12 Transformer layers with 12 attention heads) that we consider in Section 4, we include (i) a vanilla Transformers (i.e. without any pre-training) with the same parameter size as GPT-2-small; and (ii) a larger pre-trained model, GPT-2-large, that consists of 36 Transformer layers with 20 attention heads. The training of different language models follows the same procedure as described in Section 4. To measure the isotropy of the language model, we include the conicity metric [27] as well as the self-similarity

[1] https://github.com/yxuansu/SimCTG/tree/main/simctg
[2] https://pypi.org/project/simctg/
[3] https://github.com/yxuansu/SimCTG

metric (Eq. (6)). A lower conicity or self-similarity indicates the representation space of the language model better follows an isotropic distribution.

Table 1 presents the experimental results. We observe that our approach (i.e. SimCTG + contrastive search) performs the best on all evaluated models, suggesting the clear generalization ability of our approach. Another interesting finding is that, for vanilla Transformers and GPT-2-large, the model trained with MLE naturally displays a high level of isotropy. A similar phenomenon is also observed in language models from other languages, such as Chinese (see Appendix H). In such cases, our proposed contrastive search can be directly applied and yields superior performances. This further points out the huge potential of contrastive search in other much larger and stronger language models such as GPT-3 [3] and OPT [41]. We leave the rigorous investigation on the isotropic properties of different language models to our future work.

E Ablation Study on the Hyperparameters of Contrastive Search

Here, we present a detailed ablation study on the hyperparameters (i.e., k and α in Eq. (5)) of contrastive search. Specifically, we simultaneously vary the value of k and α. k is chosen from $\{5, 8, 10\}$ and α is chosen from $\{0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. For evaluation, we report the generation diversity and generation perplexity on the test set of Wikitext-103. The results are plotted in Figure 1. We see that, when k is constant, the increase of α generally increases the generation diversity and generation perplexity. When α is constant, a larger k also leads to the increased generation diversity as well as generation perplexity. Nonetheless, for different k, the overall trends are relatively the same and the value of α has more impact on the generated results. In practice, our recommended selection range of k and α are $k \in [5, 10]$ and $\alpha \in [0.5, 0.8]$, as these settings produce results that are more similar to human-written texts as judged by generation diversity and generation perplexity.

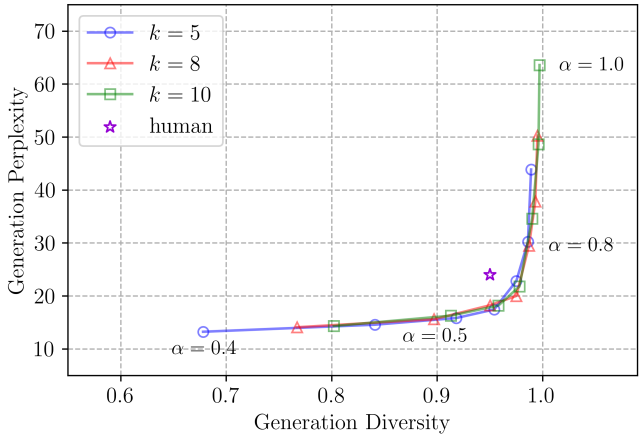

Figure 1: Ablation study on the hyperparameters of contrastive search.

F Gen-ppl Results Measured by Different Models

	greedy	beam	nucleus	contrastive	human
MLE	7.77	6.48	48.82	9.43	
Unlike.	39.02	37.38	76.22	46.03	24.86
SimCTG	8.01	6.87	47.64	**20.53**	

Table 2: The results of gen-ppl measured by the model trained with MLE.

	greedy	beam	nucleus	contrastive	human
MLE	13.18	11.67	58.01	15.94	
Unlike.	44.13	42.67	71.13	47.82	29.62
SimCTG	12.34	10.98	55.24	**23.47**	

Table 3: The results of gen-ppl measured by the model trained with Unlikelihood.

In Table 2 and 3, we show the gen-ppl (detailed in Section 4.1.2) results of different methods as measured by the model trained with MLE and Unlikelihood, respectively. As we use different models to measure gen-ppl, the results in Table 1 and 3 are slightly different from the ones in Table 1. Nontheless, we can draw the same conclusion as in Section 4.2 that SimCTG + contrastive search is

the best performing method as it obtains the generation perplexity that is closest to the human-written text.

G Human Evaluation Guidelines

Given the human-written prefix, please evaluate the system's result with respect to the following features: (1) Coherence; (2) Fluency; and (3) Informativeness. In the following, we provide some guidelines regarding how to judge the quality of the system's result in terms of different features.

G.1 Coherence

This metric measures whether the system's result is semantically and factually consistent with the human-written prefix. The definitions of different scores are:

- **[5]**: The system's result is perfectly in line with the semantic meaning defined by the prefix. And all its content is factually supported by or can be logically inferred from the prefix.
- **[4]**: The system's result is very related to the prefix but with some minor errors that does not affect its overall relevance with respect to the prefix.
- **[3]**: The system's result is, to some extent, relevant to the prefix with some errors that display minor semantic inconsistency or contradiction.
- **[2]**: At the first glance, the system's result seems to be related to the prefix. But with careful inspection, the semantic inconsistency can be easily spotted.
- **[1]**: The system's result is obviously off-the-topic or it is semantically contradicted to the content contained in the prefix.

G.2 Fluency

This metric measures the fluency of the system's result. The definitions of different scores are:

- **[5]**: The system's result is human-like, grammatically correct, and very easy to understand.
- **[4]**: Choose this score when you are hesitant between the score 3 and score 5.
- **[3]**: The system's result contains minor errors but they do not affect your understanding.
- **[2]**: Choose this score when you are hesitant between the score 1 and score 3.
- **[1]**: The system's result does not make sense and it is unreadable.

G.3 Informativeness

This metric measures the diversity, informativeness, and interestingness of the system's result. The definitions of different scores are:

- **[5]**: The system's result is very informative and contains novel content. In addition, it displays a high level of diversity and it is enjoyable to read.
- **[4]**: Choose this score when you are hesitant between the score 3 and score 5.
- **[3]**: The system's result contains some new information and it displays a certain level of diversity.
- **[2]**: Choose this score when you are hesitant between the score 1 and score 3.
- **[1]**: The system's result is dull, repetitive, and does not have new information. All its content has already been provided in the prefix.

Participant Compensation. In each experiment (i.e., open-ended text generation and open-domain dialogue generation), we hire 5 annotators to conduct the human evaluation. For every task, each annotator is paid by $400.

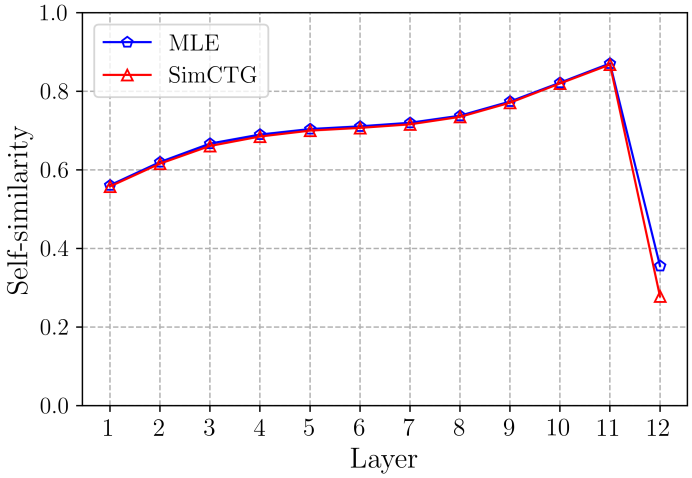

Figure 2: Layer-wise self-similarity of Chinese language models.

H Self-similarity of Chinese Language Models

We follow the same procedure as described in Section 6.1 to measure the token self-similarity of Chinese language models. Specifically, we use the test set of LCCC benchmark and compute the model's self-similarity. Figure 2 plots the layer-wise token self-similarity of the MLE and SimCTG models. We see that in all layers (including the final layer), the MLE model displays a similar self-similarity with respect to SimCTG. This observation is quite different from what we see from English language models as shown in Figure 2, where the self-similarities of SimCTG and MLE are notably different in the final layer. We conjecture that this discrepancy might come from the intrinsic property of different languages. For English, current state-of-the-art methods always represent the text into subword units, such as BPE [25], and the same subword could be over-shared by many different contexts. Thus, the representations of distinct subwords become less distinguishable which naturally leads to the anisotropy in their representations.[4] On the other hand, languages like Chinese are naturally represented by basic units, i.e., characters. Such natural unit boundary of text alleviates the over-sharing of characters in different contexts. As a result, even the vanilla MLE objective can obtain a representation space that displays a high level of isotropy.

This isotropic property of Chinese language model is particularly attractive as contrastive search can be directly applied even **without** contrastive training as shown in Table 3. In addition, we expect contrastive search could be used on off-the-shelf language models that are trained with MLE in other languages whose texts are naturally tokenized by characters (e.g., Korean and Japanese). This remains to be rigorously tested in our future work.

	MLE	**Unlikelihood**	**SimCTG**
Train FLOPs	8.08e16	8.91e16	8.20e16
Parameters	117M	117M	117M

Table 4: Training efficiency comparison.

I Training Efficiency Comparison

In this part, we compare the training efficiency of different methods (i.e., MLE, Unlikelihood, and SimCTG). To this end, we compute the total floating point operations (FLOPs) required for the training of different models on Wikitext-103. The details of training setup are provided in Section 4. Table 4 shows the results, from which we see that SimCTG is more efficient than the unlikelihood

[4]However, we should also note that, for larger English models (e.g., GPT-2-large), this conjecture not longer holds as demonstrated in Appendix D. This urges us to conduct more thorough investigations on the isotropic properties of language models across different sizes as well as different languages. We will leave these investigations to our future work.

method. Comparing with MLE, SimCTG only introduces an negligible 1.48% extra computational overhead, which further verifies the practical usage of SimCTG.

J Generated Examples on Open-domain Dialogue Generation

In Table 5, we show some generated responses of our approach (i.e., SimCTG + contrastive search) plus the reference response on examples from the test set of the Chinese LCCC benchmark. We see that, given the dialogue context, our approach is able to generate responses that are both grammatically fluent and semantically consistent with the dialogue context. These results further demonstrate the generality of our approach across different languages and tasks.

Dialogue Context	Generated Response	Reference Response
A: 老铁家好吃贾三不好吃	哈哈我觉得还可以，就是有点油	我不挑食
A:话说红海真的好看！！！准备二刷了！！！但求多排场！！！ B:我明天二刷！	哈哈哈好的！等我回学校一起吃火锅！！！	必须的！上班之后可能就没空了呢~
A:你家真有钱 B:从何说起？ A:可以买粮	好吧我承认我是土豪	你家的粮自己种的？
A:为何突然伤感?	因为没什么感觉。不知道为什么就觉得自己好孤单	因为我就要回学校了，他舍不得我
A:刺猬很可爱！以前别人送了只没养，味儿太大！！ B:是很可爱但是非常臭 A:是啊，没办法养 B:那个怎么养哦不会扎手吗	我觉得还好，就是有点臭	不会
A:杂粿太多反而摸不到主题了是吗	这就是一个故事，不知道怎么编出来的	不是，故事编不好

Table 5: Case study on the LCCC dataset.

K More Generated Examples of SimCTG + Contrastive Search

In Table 6, we provide more generated examples of SimCTG + contrastive search based on prefixes from Wikitext-103. The details of the decoding procedure are described in Section 4.

prefix	The Olmec colossal heads are at least seventeen monumental stone representations of human heads sculpted from large basalt boulders. The heads date from at least before
result	the Roman-era, and have been interpreted as representing a group of men and women that lived in the city of Arundel, the same site on which Alexander the Great's tomb was built. In addition, there is evidence to suggest that the city was part of the empire of the Persians and their conquests in Asia Minor during the 7th century BC. This may have been an intentional move to avoid confusion with the Roman city of Acre, which is located in the foothills of the mountain range near the Roman fort of Caesarea. Other evidence points to the construction ...
prefix	Although Rifenburg finished fourth among midwestern Heisman voters in 1948, he did not finish among the top eight. By comparison, Notre Dame end Leon Hart won
result	the Heisman Trophy at the end of the 1948 season after a 4 – 0 victory over Michigan on the final day of the regular season. Hart was named to the Big Ten All-Pac-10 first team, and his Heisman Trophy was awarded to Notre Dame's Bobby Bowden as the team's all-time leader in passing yards (1,000) and completions (2,065). The game ended in a 42 – 21 victory for Notre Dame, which led the nation in total offense (4,000 yards) and touchdowns ...
prefix	Kujira developed from a broad area of disturbed weather as a tropical depression on April 9 well removed from any landmasses. Shortly after development, Kuj
result	ira became extratropical, which would have caused the storm to turn to the west-northwest in the Gulf of Mexico on April 10. However, due to the proximity of the cyclone to an area of low pressure, Kujira's winds were less than 10 mph (15 km / h) at the time of landfall. By April 11, Hurricane Hunters estimated that there were 5,000 people in the area, of whom 1,000 were located in Puerto Rico. The storm weakened to a tropical storm early on April 12, and dissipated the next day. Later that day ...
prefix	The island is composed of a variety of volcanic and sedimentary rock formations from the Devonian period. At that time the Scottish landmass formed part of the Old
result	North Sea, and it was inhabited by an extensive and diverse population of sea-dwelling birds that had lived in the islands for many centuries. There is evidence of human habitation in the Isles of Scilly and St Albans, and some archaeological evidence is in the form of stone mounds dating to the 12th century. The most northerly of these is the Murchison Formation, which is located south of the island's mouth and extends to the northwest as far as Aberdeenshire. Other sites include the Isle of Man, Isle of Wight, Isle of Orkney, The Loch Ness Monster ...
prefix	A report published in the American magazine Time a fortnight later described the incident as " a genuine horror story, calculated to make the most alarming of Rhodesian doomsday
result	scenarios ". In an interview with BBC Radio 4's Tom Pertwee, he said : " There's no evidence that anyone in Rhodesia, other than the police, had done anything to provoke it ". On 20 April 2013, the government announced that it had found no evidence of involvement by Rhodesia in the attack, and would continue its investigation as soon as possible. The Independent Commission Against Corruption (ICAO) was appointed to investigate the incident in December 2013, and is the first body to do so in the UK. ICAO Director-General Mark Rowley said it was " the most ...
prefix	The American poet Ezra Pound was introduced to the group in April 1909 and found that their ideas were close to his own. In particular, Pound's studies of
result	German Expressionism (1891 – 1903) helped him develop a critical attitude towards the subject, and he had the opportunity to speak at the first meeting of the group on 25 April 1909, at New York's Waldorf-Astoria Hotel. This was the first time Pound had been invited to speak at a German Expressionist conference, and it proved to be a success. Pound became an admirer of Expressionism from the outside, and was able to attract many of its members by providing them with information and writing, such as letters to German Expressionists in Germany, as well as photographs of their works ...
prefix	The Olmec civilization developed in the lowlands of southeastern Mexico between 1500 and 400 BC. The Olmec heartland lies on the Gulf Coast of Mexico
result	, about 1,500 miles (2,500 km) west of Acapulco County, where there is a population of 1,500 people. This is the largest city in the state, and it is home to the city's most famous tourist attraction, the El Centro de las Amazons, which has been named by Mexican historian José Martínez Sánchez for its beauty and architecture. The city, along with its surroundings, were important in the development of the Maya civilization, and in later Mesoamerica and Latin America as well. Today, Mexico's ...
prefix	The discovery of a colossal head at Tres Zapotes in the nineteenth century spurred the first archaeological investigations of Olmec culture by Matthew Stirling in 1938.
result	He showed that the remains of Olmec were in good condition, and suggested that there was something to be said for the fact that they had lived long enough to be recognized as an important part of the culture and its history. Stirling's work was the first to establish what was to become known as the Peruvian-Spanish language, or Peru : the language of the people of Peru and Central America. In doing so, he drew on archaeological evidence and other sources to support his theory of the cultural origins of the Peruvian people. This is reflected in the name " Peruvian ", which stands for Peruvian ...

Table 6: More generated examples of SimCTG + contrastive search on Wikitext-103.

L Diverse Contrastive Search

In this part, we present a stochastic version of contrastive search (i.e., diverse contrastive search) which is described in Appendix A. Specifically, given the prefix with length of 32, we first generate **2** tokens using nucleus sampling with $p = 0.95$, then we use contrastive search to generate the remaining 126 tokens (i.e., 128 generated tokens in total).

prefix	Butt criticized Donald's controls in certain situations in the game, as well as the difficulty of some levels and puzzles. Buchanan also criticized the controls, calling
	Contrastive Search
result	them " unimpressive " and a " nightmare " of an experience to play with players unfamiliar with Tetris. On the other hand, his opinion was shared by other reviewers, and some were critical of the game's technical design for the Wii version of Tetris. In addition, Tintin's review included a quote from Roger Ebert, who said that Tetris was better than the original game due to its simplicity and ease of play. Ebert's comments were included in the game's DVD commentary, released on March 22, 2010. It is unclear if any of the video commentary was taken from ...
	Diverse Contrastive Search
[1]	the game " unplayable " and adding that, " It's a game in which you have to be careful not to fall into the trap of thinking too much about your opponents' strengths and weaknesses. " On the other hand, his review of the game, published in the New York Times, found little to complain about in comparison to some of the other games in the series. In particular, The Escapist's Brian O 'Toole wrote that the game was an " absolute joy ", and " one of the best-selling games of all time ". O 'Toole concluded by saying that although ...
[2]	it " a complete waste of time " and " unplayable ". On the other hand, his review of Baldur's Gate II was positive, with Buchanan commenting that, " Baldur's Gate II is an adventure game in its own right, full of fun and challenge that makes you want to go back to the first game in your life. " Buchanan felt that there were too many elements in the game for players to enjoy without some level-playing to be enjoyable at the same time. He concluded by saying that Baldur's Gate II's controls were well-balanced, and that players ...
[3]	the choice of " a simple jump button to perform a 'jump-and-a-bop' or more complex 'jump-and-a-bop' " an error and a waste of time. On the other hand, Tintin was critical of the game's design, writing that there was " too much going on " at the beginning of the game, and " not enough time " in the final cutscene for the player to make it through the game at all. He felt that the gameplay was lacking in some areas, such as the ...

Table 7: Generated results of SimCTG with diverse contrastive search.

Table 7 shows three generated results with diverse contrastive search using the same prefix as in Table 4. We see that only sampling **2** tokens at the start is enough to produce a diverse set of results. In future work, we will investigate other more sophisticated extensions of contrastive search.

References

[1] Chenxin An, Jiangtao Feng, Kai Lv, Lingpeng Kong, Xipeng Qiu, and Xuanjing Huang. Cont: Contrastive neural text generation. *arXiv preprint arXiv:2205.14690*, 2022.

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

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.

[5] Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca J. Passonneau, and Rui Zhang. Container: Few-shot named entity recognition via contrastive learning. *CoRR*, abs/2109.07589, 2021.

[6] Angela Fan, Mike Lewis, and Yann N. Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics, 2018.

[7] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *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 6894–6910. Association for Computational Linguistics, 2021.

[8] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021.

[10] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. Opennmt: Open-source toolkit for neural machine translation. In Mohit Bansal and Heng Ji, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, System Demonstrations*, pages 67–72. Association for Computational Linguistics, 2017.

[11] Wouter Kool, Herke van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *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 3499–3508. PMLR, 2019.

[12] Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. Rankgen: Improving text generation with large ranking models. *arXiv preprint arXiv:2205.09726*, 2022.

[13] Tian Lan, Deng Cai, Yan Wang, Yixuan Su, Xian-Ling Mao, and Heyan Huang. Exploring dense retrieval for dialogue response selection. *CoRR*, abs/2110.06612, 2021.

[14] Fangyu Liu, Ivan Vulic, Anna Korhonen, and Nigel Collier. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *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 1442–1459. Association for Computational Linguistics, 2021.

[15] Yixin Liu and Pengfei Liu. Simcls: A simple framework for contrastive learning of abstractive summarization. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 1065–1072. Association for Computational Linguistics, 2021.

[16] Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. Neurologic a*esque decoding: Constrained text generation with lookahead heuristics. *CoRR*, abs/2112.08726, 2021.

[17] Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Neurologic decoding: (un)supervised neural text generation with predicate logic constraints. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pages 4288–4299. Association for Computational Linguistics, 2021.

[18] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics, 2015.

[19] Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Typical decoding for natural language generation, 2022.

[20] Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. COCO-LM: correcting and contrasting text sequences for language model pretraining. *CoRR*, abs/2102.08473, 2021.

[21] Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. *CoRR*, abs/2110.08173, 2021.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[23] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[24] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In Regina Barzilay and Min-Yen Kan, editors, *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 1073–1083. Association for Computational Linguistics, 2017.

[25] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics, 2016.

[26] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 1134–1141. IEEE, 2018.

[27] Aditya Sharma, Partha Talukdar, et al. Towards understanding the geometry of knowledge graph embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 122–131, 2018.

[28] Yixuan Su, Deng Cai, Yan Wang, David Vandyke, Simon Baker, Piji Li, and Nigel Collier. Non-autoregressive text generation with pre-trained language models. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 234–243. Association for Computational Linguistics, 2021.

[29] Yixuan Su, Deng Cai, Qingyu Zhou, Zibo Lin, Simon Baker, Yunbo Cao, Shuming Shi, Nigel Collier, and Yan Wang. Dialogue response selection with hierarchical curriculum learning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1740–1751. Association for Computational Linguistics, 2021.

[30] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*, 2022.

[31] Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi, and Nigel Collier. Tacl: Improving BERT pre-training with token-aware contrastive learning. *CoRR*, abs/2111.04198, 2021.

[32] Yixuan Su, Zaiqiao Meng, Simon Baker, and Nigel Collier. Few-shot table-to-text generation with prototype memory. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 910–917. Association for Computational Linguistics, 2021.

[33] Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. Multi-task pre-training for plug-and-play task-oriented dialogue system. *CoRR*, abs/2109.14739, 2021.

[34] Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. Plan-then-generate: Controlled data-to-text generation via planning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 895–909. Association for Computational Linguistics, 2021.

[35] Yixuan Su, Yan Wang, Deng Cai, Simon Baker, Anna Korhonen, and Nigel Collier. PROTOTYPE-TO-STYLE: dialogue generation with style-aware editing on retrieval memory. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:2152–2161, 2021.

[36] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112, 2014.

[37] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[38] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[39] Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. Challenges in data-to-document generation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2253–2263. Association for Computational Linguistics, 2017.

[40] Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir R. Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *CoRR*, abs/2201.05966, 2022.

[41] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.